

NUMERICAL METHODS IN EXTREMAL PROBLEMS

B. N. Pshenichny and Yu. M. Danilin

**MIR
PUBLISHERS
MOSCOW**

NUMERICAL METHODS IN EXTREMAL PROBLEMS

B.N. Pshenichny and Yu. M. Danilin

The book describes methods and algorithms for numerical solution of problems of finding extrema of functions and functionals met with in mathematical programming, economics, optimal control theory and other fields of science and practice.

Special attention is paid to algorithms with a fast rate of convergence and implementable on computers. Methods of unconstrained and constrained minimization of functions of independent variables are discussed. The book will be useful to specialists in mathematical programming, computational mathematics and optimal control theory and to broad circles of students and engineers, who in their practical work have to solve problems of function minimization.

NUMERICAL METHODS IN EXTREMAL PROBLEMS

B. N. Pshenichny
and
Yu. M. Danilin

Б. Н. Пшеничный

Ю. М. Данилин

ЧИСЛЕННЫЕ

МЕТОДЫ

В ЭКСТРЕМАЛЬНЫХ

ЗАДАЧАХ

ИЗДАТЕЛЬСТВО «НАУКА»

ГЛАВНАЯ РЕДАКЦИЯ

ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ

МОСКВА



NUMERICAL METHODS IN EXTREMAL PROBLEMS

B. N. Pshenichny and Yu. M. Danilin

Translated from the Russian
by V. Zhitomirsky, D. Sc. (Eng.)

MIR
PUBLISHERS
MOSCOW

First published 1978

**Revised from the 1975
Russian edition**

На английском языке

**© Главная редакция физико-математической литературы
издательства «Наука», 1975**

© English translation, Mir Publishers, 1978

CONTENTS

PREFACE 9

CHAPTER I. INTRODUCTION TO THE THEORY OF MATHEMATICAL PROGRAMMING 12

1. CONVEX SETS 12

Definition. Separation Theorem. Convex Cones. Strictly and Strongly Convex Sets.

2. CONVEX FUNCTIONS 17

Definition. Basic Properties. Differential Properties. Strictly and Strongly Convex Functions. Concave Functions.

3. CONVEX PROGRAMMING 24

Formulation of the Problem. Basic Properties. Necessary Conditions for a Minimum. The Kuhn-Tucker Theorem. Dual Problem. Problem of Linear Programming. Problem of Quadratic Programming.

4. NECESSARY CONDITIONS FOR A MINIMUM 32

Basic Definitions. Necessary Conditions for a Minimum. Minimax Problem. Necessary Conditions of the Second Order.

5. SOME ADDITIONAL INFORMATION 41

Bibliographic Notes 42

CHAPTER II. METHODS OF UNCONSTRAINED FUNCTION MINIMIZATION 44

1. GRADIENT METHODS 45

Method of Steepest Descent. Variants of the Method. Other Gradient Methods. Qualitative Analysis of the Methods.

2. NEWTON'S METHOD WITH STEP ADJUSTMENT 58

Construction of the Method. Theorems about Properties of the Method. Modifications of the Generalized Newton Method. Discussion of the Properties of Newton's Method.

3. METHODS OF DUAL DIRECTIONS 67

Considerations on the Choice of Schemes of the Methods. Substantiation of the Methods. Construction of Various Algorithms. Determining Vector p_k . The Initial Stage of the Process. Minimization of Quadratic Form. Discussion of Properties of the Methods.

4. METHODS OF CONJUGATE DIRECTIONS. MINIMIZATION OF QUADRATIC FUNCTIONS 82

Conjugate Directions and Their Properties. Construction of the Methods. General Properties of the Methods. Concrete Algorithms. Minimization of a Convex Quadratic Function. Discussion of Results.

5. METHODS OF CONJUGATE DIRECTIONS. MINIMIZATION OF ARBITRARY FUNCTIONS 103

Considerations about the Applicability of the Methods. Theorem on Convergence of the Methods. Study of Properties of Different Algorithms. Further Study of the Rate of Convergence. Discussion of Results.

6. METHODS WITHOUT CALCULATING DERIVATIVES 129

Introductory Remarks. Constructing Methods of Dual Directions. Remarks on the Implementation of Methods of Dual Directions. Methods of Conjugate Directions. Discussion of Results.

Bibliographic Notes 145

CHAPTER III. METHODS OF CONSTRAINED FUNCTION MINIMIZATION 146**1. PROBLEM OF QUADRATIC PROGRAMMING 146**

Operators of Projection. Minimization of a Quadratic Function in a Subspace. Algorithm of General Problem of Quadratic

Programming. Computational Aspects. Problem of Quadratic Programming with Simple Constraints.

2. METHOD OF FEASIBLE DIRECTIONS 162

Method of Choosing Feasible Directions. Algorithm of Method of Feasible Directions. Substantiation of Convergence of the Algorithm. Construction of the Initial Approximation.

3. METHOD OF CONDITIONAL GRADIENT AND NEWTON'S METHOD 170

Rule for Choosing the Step Length. Description of the Algorithm. Substantiation of Convergence of the Algorithm and Estimation of Its Rate of Convergence. Estimate of Convergence for a Strongly Convex Region. Newton's Method with Step Adjustment. Properties of Newton's Method.

4. CUTTING HYPERPLANE METHOD 184

Algorithm. Computational Aspects. Concluding Remarks.

5. LINEARIZATION METHOD 188

Basic Assumptions. Formulation of the Algorithm. Convergence of the Algorithm. Computational Aspects. Some Generalizations. Problem of Linear Programming. Local Estimate of the Rate of Convergence.

6. LINEARIZATION METHOD: SOLVING SYSTEMS OF EQUALITIES AND INEQUALITIES AND FINDING THE MINIMAX 211

Systems of Equalities and Inequalities. Convergence of the Algorithm. Remarks. Sufficient Conditions of Convergence. Solving the Problem of Finding the Minimax.

7. LOCAL ACCELERATION OF CONVERGENCE 224

Formulation of the Problem. Basic Formulas. Algorithm. Computational Aspects. Application to the Problem of Mathematical Programming. Minimization Problem with Equality Constraints.

8. METHOD OF PENALTY FUNCTIONS 235

Substantiation of the Penalty Function Method. Convex Programming. Computational Aspects. Fiacco and McCormick Method.

9. PROJECTION METHODS WITH RESTORATION OF TIES 244

Construction of the Methods. Methods of the First Order. Method of the Second Order. Minimization Methods of Higher Effectiveness. On the Solving of the General Problem of Mathematical Programming. Conclusive Remarks.

Bibliographic Notes 257

APPENDIX. COMPUTATIONAL SCHEMES OF THE MAIN ALGORITHMS 259

LITERATURE 265

INDEX 271

PREFACE

Computational methods of solving extremal problems developed very intensively in recent years.

The lists of the literature on these subjects contain at present hundreds of items. This interest in the development of computational methods is not casual. It reflects the important role played by the finding of extrema in diverse applied problems. The problem of an effective minimization of a function with different constraints on the variables is the subject matter of this book.

It should be stressed from the very beginning that recent years have brought changes in the requirements to be met by new computational algorithms. Some ten or fifteen years ago any new algorithm for solving a minimization problem was noticed with interest, but now only the construction of a new algorithm is insufficient. It is now necessary to show in what respect it is better than the existing ones. Thus there arises the problem of comparing the effectiveness of different algorithms. Unfortunately this problem has no simple solution. This is due to the necessity of choosing a criterion of effectiveness and the criteria may be diverse. For instance, we can take as a criterion of effectiveness the accuracy of the result obtained, the time required for computing, the necessary storing capacity of the computer, etc. Also it is often necessary to use rather contradictory criteria in estimating an algorithm.

In selecting algorithms to be included in this book, the authors based their choice on the criterion of accuracy of the result and the rate of convergence of the iterative process. However, even with this limiting condition it is not possible to order all the algorithms in one and only one way and tell which of them is better or worse than another. The reason is that the estimate of the rate of convergence is not made for a particular problem, rather it is applied to a class of problems. Therefore an algorithm which is poor as applied to a broad class of problems can prove effective on a narrower one. This makes it necessary for the calculator to keep a large reserve of algorithms and to apply them depending on the problem to be solved.

It is important to know what ensures a fast rate of convergence of the algorithm. In practice, even the calculating of the first derivative of a function quite often involves certain difficulties; these become insurmountable when trying to calculate the second derivative.

Therefore special stress is laid on the description of the algorithms that require the finding only of the first derivative or only of the value of the function.

In describing the computational methods we consider only the finite dimensional case. This is due to two reasons. First, in using a computer for calculations, the problem is to be approximated anyway by a finite dimensional one. Secondly, most of the known algorithms are comparatively simply generalized for the minimization of functionals without essential changes. This approach made it possible to make the book easily understood by a broad circle of readers, since in order to grasp most of the results described only a knowledge of the principles of mathematical analysis and linear algebra is required.

To avoid the necessity of frequent cross-referencing, not many references are given in the text. Short bibliographic notes follow some of the chapters. The authors did not attempt to comprise all the literature on the questions treated, this being simply impossible because of its vastness. This is why the list of literature given at the end of the book includes only papers and monographs directly used in writing this book.

It should be noted that the authors have not discussed the methods of solving a broad and important class of noncorrect extremal problems, which are treated in the works of A. N. Tikhonov and his followers. The authors have but slightly touched the solving of optimal control problems. These problems have been studied from various points of view and the methods for their solution are given in N. N. Moiseev's monograph *Numerical Methods in the Theory of Optimal Systems*.

The algorithms set forth below are iterative in character. This means that we can construct a finite or infinite sequence of points x_k , $k = 0, 1 \dots$ which is said to converge to the solving of a minimization problem.

The points of the sequence are related by the equation

$$x_{k+1} = x_k + \alpha_k p_k$$

where p_k is the vector of shift from point x_k and α_k is a step along the direction of p_k . Therefore the description of any of the algorithms given below consists in establishing the method of choosing the vector p_k and the length of the step α_k . It should be noted that the method of choosing the vector p_k determines the general rate of convergence of the process and the method of choosing α_k has an important influence on the amount of calculations at each iteration. Therefore the authors' aim was to give in all cases of choosing α_k a method, such that the required value of α_k could be found after a finite number of iterations without affecting the general rate of convergence.

Let us briefly review the estimates of the rate of convergence, which are in most cases used in this book.

We say that a sequence $\{x_k\}$ converges to point x_* at a *linear rate* or at the *rate of geometrical progression* (with the ratio q) if from a certain k the inequality $\|x_{k+1} - x_*\| \leq q \|x_k - x_*\|$ where $0 < q < 1$, is satisfied. If the inequality $\|x_{k+1} - x_*\| \leq q_k \|x_k - x_*\|$ is satisfied, where $q_k \rightarrow 0$ with $k \rightarrow \infty$, we say that the *rate of convergence of the sequence $\{x_k\}$ is superlinear*, or *faster than the rate of convergence of any geometric progression*. If $q_k \leq C \|x_k - x_*\| \rightarrow 0$, then $\|x_{k+1} - x_*\| \leq C \|x_k - x_*\|^2$. This estimate is a characteristic of the *quadratic rate of convergence*.

The above estimates will occur in this book also in several other equivalent forms.

Some remarks on the notations used.

As mentioned before, the subject is treated for the case of an n -dimensional vector space which will be denoted by E^n . The vectors will be denoted by lower-case letters x, y, z , etc. and their components by using superscripts so that x^i is the i -th component of vector x . The subscripts denote the elements of a sequence. Matrices are denoted by capital letters A, B, C etc. An asterisk as upper index denotes transposition, i.e. A^* is the transposed matrix A . As a rule vector x means a column-vector so that x^* denotes a row-vector. The scalar product of two vectors is denoted by (x, y) , i.e.

$$(x, y) = \sum_{i=1}^n x^i y^i.$$

The norm of the vector is understood to be its Euclidean norm, unless otherwise specified:

$$\|x\| = \sqrt{(x, x)}.$$

In conclusion, the authors express their sincere gratitude to G. E. Lybarskaya, L. A. Sobolenko, E. I. Boguslavskaya and V. M. Panin for the invaluable assistance in preparing this book.

Chapter I (except Sec. 5 and partly Sec. 2) and Chap. III (except Sec. 9 and partly Sec. 3) have been written by B. N. Pshenichny.

Chapter II, the third and the fourth subsections of Sec. 2 and the fifth and sixth subsections of Sec. 3, and Sec. 9 of Chap. III have been written by Yu. M. Danilin.

CHAPTER I

INTRODUCTION TO THE THEORY OF MATHEMATICAL PROGRAMMING

This chapter describes some facts from the theory of convex sets and the necessary conditions of the extrema; these facts are necessary for understanding the matter set forth in subsequent chapters.

1. CONVEX SETS

In this section we consider the basic properties of convex sets in an n -dimensional Euclidean space.

Definition. Separation Theorem

Definition 1.1. *A set of points X in E^n is called convex if together with any $x_1, x_2 \in X$ it contains also all points of the form:*

$$x = \lambda x_1 + (1 - \lambda) x_2, \quad 0 \leq \lambda \leq 1.$$

In geometrical terms this means that if the end points of a segment belong to a convex set X then the whole segment belongs to the set too.

Lemma 1.1. *The following statements hold:*

(1) *The intersection of any number of convex sets is convex.'*

(2) *If $x_i \in X$, $i = 1, \dots, m$, then with any λ_i , $i = 1, \dots, m$*

such that $\sum_{i=1}^m \lambda_i = 1$, $\lambda_i \geq 0$, a point $x = \sum_{i=1}^m \lambda_i x_i$ belongs to X .

The following theorem and its corollaries are the basic tools using which it is possible to obtain results characterising various properties of convex sets.

Theorem 1.1. *Let X be a convex set, and $\bar{X}$ its closure. If point x_0 does not belong to $\bar{X}$, then there exist a vector $a \in E^n$, $a \neq 0$, and a number $\varepsilon > 0$ such that for all $x \in X$*

$$(a, x) \leq (a, x_0) - \varepsilon.$$

Proof. $\bar{X}$ is a closed set, by definition. Let us show that it is convex. Indeed, if $x \in \bar{X}$, then there is a sequence $\{x_k\}$, $k = 1, \dots$, such that $x_k \in X$, $x_k \rightarrow x$. Now let $x, y \in \bar{X}$, $0 \leq \lambda \leq 1$. Let us prove that $\lambda x + (1 - \lambda)y \in \bar{X}$. Since X is a convex set, it follows from $x_k, y_k \in X$, $x_k \rightarrow x$, $y_k \rightarrow y$ that

$$\begin{aligned}\lambda x_k + (1 - \lambda) y_k &\in X, \\ \lambda x_k + (1 - \lambda) y_k &\rightarrow \lambda x + (1 - \lambda) y.\end{aligned}$$

This means that $\lambda x + (1 - \lambda) y \in \bar{X}$, i.e. $\bar{X}$ is convex.

Let us take a point $y \in \bar{X}$ whose distance from x_0 is the least, i.e.

$$\|x - x_0\| \geq \|y - x_0\|, \quad x \in \bar{X}.$$

Since $\bar{X}$ is convex for all $x \in X$ and $0 \leq \lambda \leq 1$, we have

$$\lambda x + (1 - \lambda) y = y + \lambda (x - y) \in X.$$

Therefore

$$\begin{aligned}\|\lambda x + (1 - \lambda) y - x_0\|^2 &= \|y - x_0 + \lambda (x - y)\|^2 \\ &= (y - x_0 + \lambda (x - y), y - x_0 + \lambda (x - y)) \\ &= (y - x_0, y - x_0) + 2\lambda (y - x_0, x - y) + \lambda^2 (x - y, x - y) \\ &= \|y - x_0\|^2 + 2\lambda (y - x_0, x - y) + \lambda^2 \|x - y\|^2 \geq \|y - x_0\|^2.\end{aligned}$$

The last inequality holds for any λ , varying between zero and unity. Simplifying it we obtain

$$2(y - x_0, x - y) + \lambda \|x - y\|^2 \geq 0;$$

hence with $\lambda = 0$

$$(y - x_0, x - y) \geq 0.$$

Let $a = x_0 - y$. The last inequality can then be written in the form $(a, x) \leq (a, y)$. But

$$(a, y) = (a, x_0) - (a, x_0 - y) = (a, x_0) - \|a\|^2.$$

Setting $\varepsilon = \|a\|^2$, we finally obtain

$$(a, x) \leq (a, x_0) - \varepsilon.$$

This inequality holds for any $x \in \bar{X}$. Besides $\varepsilon > 0$ as $x_0 \notin \bar{X}$ and consequently $y \neq x_0$. Therefore

$$\varepsilon = \|a\|^2 = \|x_0 - y\|^2 > 0.$$

Q.E.D.

Remark. In proving theorem 1.1. we have proved at the same time that the closure of a convex set is convex too. As a simple exercise the reader can prove that the set of interior points of a convex set is convex too.

Corollary 1.1. *Let X be a convex set and x_0 the frontier point of X . Then there is a vector $a \neq 0$ such that*

$$(a, x) \leq (a, x_0), \quad x \in X.$$

Corollary 1.2. *If X and Y are convex sets that do not intersect, then there is a vector $a \neq 0$ such that*

$$(a, x) \leq (a, y), \quad x \in X, y \in Y.$$

Corollary 1.3. *If X and Y are closed convex sets which do not intersect and one of them is bounded, then there exist a vector $a \neq 0$ and a number $\varepsilon > 0$ such that*

$$(a, x) \leq (a, y) - \varepsilon, \quad x \in X, y \in Y.$$

Convex Cones

Definition 1.2. *A set K is called a convex cone if the set is convex and together with every point $x \in K$ it contains all points λx with $\lambda > 0$.*

It is clear that if $x, y \in K$ then $x + y \in K$. In fact, since K is a convex set, point $\frac{1}{2}x + \frac{1}{2}y$ belongs to K . But

$$x + y = 2 \left(\frac{1}{2}x + \frac{1}{2}y \right),$$

whence $x + y \in K$ by the definition of a cone. The most important properties of cones are formulated in terms which establish the relation between the original cone and the cone that is its conjugate or dual.

Definition 1.3. *Let K be a convex cone. The set of all vectors $y \in E^n$ satisfying for any $x \in K$ the inequality $(x, y) \geq 0$ is called a conjugate cone and denoted by K^* .*

An elementary check shows that K^* is also a convex cone.

Lemma 1.2. *K^* is a closed convex cone.*

Lemma 1.3. *Let K be a convex cone. Then $x_0 \in \bar{K}$ if and only if $(x_0, y) \geq 0$ for all $y \in K^*$. If K is closed, then*

$$(K^*)^* = K.$$

Proof. It is evident that if $x_0 \in \bar{K}$, then $(x_0, y) \geq 0$ for all $y \in K^*$. Suppose it is false. Let $(x_0, y) \geq 0$ for any $y \in K^*$, but $x_0 \notin \bar{K}$.

Since K is a closed convex set and using theorem 1.1, we can assert that there is a vector a such that

$$(a, x_0) \leq (a, x) - \varepsilon, \quad x \in \bar{K}.$$

Now a closed cone K always contains point 0. Therefore in particular

$$(a, x_0) \leq -\varepsilon. \quad (1.1)$$

On the other hand

$$(a, x) \geq 0, \quad x \in \bar{K}. \quad (1.2)$$

Indeed, if for a certain $x_1 \in K$ $(a, x_1) < 0$, then since $\lambda x_1 \in K$ with $\lambda > 0$

$$(a, x_0) \leq \lambda (a, x_1) - \varepsilon$$

and the last inequality must be valid for any λ ; this is impossible if $(a, x_1) < 0$. Thus (1.2) is valid and consequently $a \in K^*$. Then $(a, x_0) \geq 0$ and this contradicts (1.1). This proves the first part of the lemma.

Let us now prove its second part. If $x \in K$, then $(x, y) \geq 0$ for all $y \in K^*$, by definition, and therefore $x \in (K^*)^*$, $K \subset (K^*)^*$. Conversely, by definition, $x \in (K^*)^*$ if and only if $(x, y) \geq 0$ with any $y \in K^*$. However, it was proved above that in this case $x \in K$, i.e. $(K^*)^* \subset K$. Thus $(K^*)^* = K$. Q.E.D.

Polyhedral cones are an important class of cones encountered in the theory of linear programming.

Definition 1.4. A cone K is called *polyhedral* if there exists a finite set of n -dimensional vectors a_i , $i = 1, \dots, m$ such that with $x \in K$ the expansion

$$x = \sum_{i=1}^m \lambda_i a_i, \quad \lambda_i \geq 0, \quad i = 1, \dots, m \quad (1.3)$$

is valid and conversely (1.3) implies that $x \in K$.

Thus a polyhedral cone K is a set of points which can be represented in the form (1.3). A given point $x \in K$ in the form (1.3), speaking generally, is represented not uniquely.

Lemma 1.4. Let $x \in K$, K being a polyhedral cone. Then there is such an expansion of x in vectors a_i with nonnegative coefficients λ_i , that the number of indices i for which $\lambda_i > 0$ does not exceed n , the number of dimensions of the space; the vectors a_i corresponding to nonzero λ_i are linearly independent.

Proof. Let $x \in K$, i.e. $x = \sum_{i=1}^m \lambda_i a_i$, and $\mathcal{J}$ be the set of those indices i such that $\lambda_i > 0$. Suppose that the number of elements in $\mathcal{J}$ is greater than n , or does not exceed n , but the vectors a_i , $i \in \mathcal{J}$, are

linearly dependent. Since more than n linearly independent vectors cannot exist in an n -dimensional space, there are coefficients α_i , not all zero, such that $\sum_{i \in \mathcal{J}} \alpha_i a_i = 0$. Besides, by definition of $\mathcal{J}$, $\lambda_i = 0$ if $i \notin \mathcal{J}$ and so

$$x = \sum_{i \in \mathcal{J}} \lambda_i a_i, \quad \lambda_i > 0, \quad i \in \mathcal{J}.$$

Subtracting from this relation the preceding one multiplied by ε , we obtain

$$x = \sum_{i \in \mathcal{J}} (\lambda_i - \varepsilon \alpha_i) a_i.$$

Without loss of generality we can take that $\alpha_i > 0$, for some $i \in \mathcal{J}$. Setting $\varepsilon_0 = \min_{i \in \mathcal{J}, \alpha_i > 0} \frac{\lambda_i}{\alpha_i}$ and $\bar{\lambda}_i = \lambda_i - \varepsilon_0 \alpha_i$, we have

$$x = \sum_{i \in \mathcal{J}} \bar{\lambda}_i a_i$$

where $\bar{\lambda}_i \geq 0$ and for one i at least $\bar{\lambda}_i = 0$.

Thus we have obtained an expansion of x in vectors a_i with non-negative coefficients; however the number of strictly positive coefficients has been diminished.

This process can now be applied further until the number of non-zero coefficients becomes less than n or equal to n and vectors a_i for which $\lambda_i > 0$ become linearly independent. Since we have a process of diminishing a whole number, this process obviously cannot be continued infinitely and after a certain number of steps we shall get an expansion which satisfies the conditions of our lemma.

Lemma 1.5. *A polyhedral cone is closed.*

Lemma 1.6. *Let the cone K be defined by a system of linear inequalities*

$$(a_i, x) \geq 0, \quad i = 1, \dots, m$$

where $a_i \in E^n$. Then the conjugate cone K^* is a polyhedral cone and consists of points y , which can be presented in the form

$$y = \sum_{i=1}^m \lambda_i a_i, \quad \lambda_i \geq 0, \quad i = 1, \dots, m.$$

Proof. Let us consider the cone

$$\tilde{K} = \left\{ y : y = \sum_{i=1}^m \lambda_i a_i, \quad \lambda_i \geq 0, \quad i = 1, \dots, m \right\}.$$

By definition, $\tilde{K}^*$ is a set of points x , for which $(x, y) \geq 0$, $y \in \tilde{K}$, i.e. $(x, \sum_{i=1}^m \lambda_i a_i) \geq 0$ for all $\lambda_i \geq 0$. Then

$$(x, \sum_{i=1}^m \lambda_i a_i) = \sum_{i=1}^m \lambda_i (x, a_i) \geq 0.$$

The last inequality can obviously be valid for any $\lambda_i \geq 0$ only if $(a_i, x) \geq 0$, $i = 1, \dots, m$, i.e. if $x \in K$. Thus $\tilde{K}^* = K$. Since $\tilde{K}$ is a polyhedral cone, it is closed and by lemma 1.3 $(\tilde{K}^*)^* = \tilde{K}$. Thus $K^* = \tilde{K}$. Q.E.D.

Remark. The lemma proved above is known as the Farkas-Minkowski lemma and is used as the basic tool for obtaining the necessary conditions for extrema.

Strictly and Strongly Convex Sets

Definition 1.5. A set $X \subset E^n$ is called strictly convex if for any $x_1, x_2 \in X$, $x_1 \neq x_2$ all points of the form

$$\lambda x_1 + (1 - \lambda) x_2, \quad 0 < \lambda < 1$$

are internal points of this set.

Definition 1.6. A set $X \subset E^n$ is called strongly convex if there is a constant $\gamma > 0$ such that any point

$$\frac{x_1 + x_2}{2} + y \in X$$

if $x_1, x_2 \in X$ and $\|y\| \leq \gamma \|x_2 - x_1\|^2$.

It is easily ascertained that a strongly convex set is also strictly convex (but not the converse).

2. CONVEX FUNCTIONS

Convex functions have a number of important properties and constitute one of the main objects of study in the theory of mathematical programming. The problem of convex programming which is the most investigated one for extrema is formulated in terms of convex sets. However convex functions play a decisive role in the general nonlinear problem too, since the sufficiently general and comprehensive necessary conditions of extrema can be formulated only for the case where the derivatives of the functions in the direction at the given point are convex functions.

We shall mainly study convex functions defined over the whole space so that the value of any given convex function is finite at each point $x \in E^n$. From the viewpoint of general theory it is sometimes expedient to consider convex functions which can at some points

take the value of $+\infty$. In what follows however such functions will occur only in studying dual problems of convex programming. Therefore in all statements of this section, unless otherwise specified, we suppose that the convex function under consideration is defined over the whole space E^n and takes finite values.

Definition. Basic Properties

Definition 2.1. A function $f(x)$ defined for all $x \in E^n$ is called convex if for any x_1, x_2 and $\lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 = 1$,

$$f(\lambda_1 x_1 + \lambda_2 x_2) \leq \lambda_1 f(x_1) + \lambda_2 f(x_2).$$

Remark. If $f(x) = +\infty$ for some x , the definition remains valid.

Lemma 2.1. Let $f_1(x)$ and $f_2(x)$ be convex functions and c_1, c_2 non-negative numbers. Then

$$f(x) = c_1 f_1(x) + c_2 f_2(x)$$

is a convex function too.

Lemma 2.2. Let $f_i(x), i = 1, \dots, m$ be convex functions. Then $f(x) = \max_{1 \leq i \leq m} f_i(x)$ is also a convex function.

Lemma 2.3. If $f(x)$ is a convex function, then we have

$$\begin{aligned} & f(\lambda_1 x_1 + \lambda_2 x_2 + \dots + \lambda_m x_m) \\ & \leq \lambda_1 f(x_1) + \lambda_2 f(x_2) + \dots + \lambda_m f(x_m) \end{aligned}$$

for any nonnegative λ_i , which satisfy the condition

$$\lambda_1 + \dots + \lambda_m = 1.$$

Proof. With $m = 2$ this statement follows from the definition of a convex function. Suppose the lemma has been proved for $m \leq k$. Let us show that the statement is valid for $m = k + 1$. Let $\lambda_i \geq 0, i = 1, \dots, k + 1, \lambda_1 + \dots + \lambda_{k+1} = 1$. Evidently one can consider all λ_i to be strictly greater than zero; otherwise we should have the case where the above inequality is satisfied by hypothesis.

Thus $\lambda_{k+1} > 0$ and $1 - \lambda_{k+1} = \lambda_1 + \dots + \lambda_k > 0$.

From the definition of a convex function we have

$$\begin{aligned} & f(\lambda_1 x_1 + \dots + \lambda_k x_k + \lambda_{k+1} x_{k+1}) \\ & \leq (1 - \lambda_{k+1}) f\left(\frac{\lambda_1}{1 - \lambda_{k+1}} x_1 + \dots + \frac{\lambda_k}{1 - \lambda_{k+1}} x_k\right) + \lambda_{k+1} f(x_{k+1}). \end{aligned} \quad (2.1)$$

But by induction

$$\begin{aligned} & f\left(\frac{\lambda_1}{1 - \lambda_{k+1}} x_1 + \dots + \frac{\lambda_k}{1 - \lambda_{k+1}} x_k\right) \\ & \leq \frac{\lambda_1}{1 - \lambda_{k+1}} f(x_1) + \dots + \frac{\lambda_k}{1 - \lambda_{k+1}} f(x_k) \end{aligned} \quad (2.2)$$

since

$$\frac{\lambda_1}{1-\lambda_{k+1}} + \dots + \frac{\lambda_k}{1-\lambda_{k+1}} = 1.$$

Comparing (2.1) and (2.2) we obtain the required result.

The lemma has been proved using the principle of mathematical induction.

Lemma 2.4. *The function $f(x)$ is convex if and only if for any x and $p \in E^n$ the function of the one-dimensional variable t*

$$\varphi_{x,p}(t) \equiv f(x + tp) \quad (2.3)$$

is a convex function.

Differential Properties

Let $f(x)$ be a convex differentiable function whose continuous gradient is $f'(x)$.

Lemma 2.5. *The following statements are equivalent:*

(1) $f(x)$ is a convex function.

(2) $f(x_2) - f(x_1) \geq (f'(x_1), x_2 - x_1)$ for any $x_1, x_2 \in E^n$.

(3) $(f'(x + \lambda p), p)$ is a nondecreasing function of λ .

If $f(x)$ is a twice continuously differentiable function, then

(4) $f''(x)$, the matrix of second derivatives, is positive definite, i.e. $(f''(x)p, p) \geq 0$ for any $x, p \in E^n$.

Proof. Note first of all that if

$$\varphi_{x,p}(\lambda) = f(x + \lambda p),$$

then as shown above $\varphi_{x,p}(\lambda)$ is a convex function and

$$\varphi'_{x,p}(\lambda) = (f'(x + \lambda p), p), \quad \varphi''_{x,p}(\lambda) = (p, f''(x + \lambda p)p). \quad (2.4)$$

Let us show that statement (2) follows from statement (1). In fact, since

$$f((1-\lambda)x_1 + \lambda x_2) \leq (1-\lambda)f(x_1) + \lambda f(x_2), \quad 0 < \lambda < 1$$

we have

$$\frac{f(x_1 + \lambda(x_2 - x_1)) - f(x_1)}{\lambda} \leq f(x_2) - f(x_1).$$

Taking the limit with $\lambda \rightarrow 0$ we obtain

$$(f'(x_1), x_2 - x_1) \leq f(x_2) - f(x_1).$$

Thus statement (2) follows from (1) or shortly (1) $\rightarrow$ (2).

Let us show that (2) $\rightarrow$ (3). From statement (2) we have for $\varphi_{x,p}(\lambda)$:

$$\begin{aligned} \varphi'_{x,p}(\lambda_1)(\lambda_2 - \lambda_1) &\leq \varphi_{x,p}(\lambda_2) - \varphi_{x,p}(\lambda_1), \\ \varphi'_{x,p}(\lambda_2)(\lambda_1 - \lambda_2) &\leq \varphi_{x,p}(\lambda_1) - \varphi_{x,p}(\lambda_2). \end{aligned}$$

The two inequalities with $\lambda_2 > \lambda_1$ give

$$\varphi'_{x,p}(\lambda_1) \leq \frac{\varphi_{x,p}(\lambda_2) - \varphi_{x,p}(\lambda_1)}{\lambda_2 - \lambda_1} \leq \varphi'_{x,p}(\lambda_2),$$

i.e. $(f'(x + \lambda_1 p), p) \leq (f'(x + \lambda_2 p), p)$. Q.E.D.

(3) $\rightarrow$ (1). Let $(f'(x + \lambda p), p)$ be a nondecreasing function of λ . Then $\varphi'_{x,p}(\lambda_1) \leq \varphi'_{x,p}(\lambda_2)$ with $\lambda_2 \geq \lambda_1$.

If $0 < \mu < 1$, then

$$\begin{aligned} 0 &\leq \mu(\lambda_2 - \lambda_1) \int_0^1 [\varphi'_{x,p}(\lambda_1 + \tau(\lambda_2 - \lambda_1)) - \varphi'_{x,p}(\lambda_1 + \tau\mu(\lambda_2 - \lambda_1))] d\tau \\ &= (1 - \mu)\varphi_{x,p}(\lambda_1) + \mu\varphi_{x,p}(\lambda_2) - \varphi_{x,p}((1 - \mu)\lambda_1 + \mu\lambda_2), \end{aligned}$$

i.e. $\varphi_{x,p}(\lambda)$ is a convex function of λ . Then as follows from lemma 2.4, $f(x)$ is a convex function.

(3) $\rightarrow$ (4). Since $\varphi'_{x,p}(\lambda) = (f'(x + \lambda p), p)$ is a nondecreasing function, $\varphi''_{x,p}(\lambda) \geq 0$, i.e.

$$(p, f''(x + \lambda p)p) \geq 0. \quad (2.5)$$

Hence the matrix $f''(x)$ is positive definite.

(4) $\rightarrow$ (3). Conversely, if (2.5) is satisfied, then $\varphi''_{x,p}(\lambda)$ is non-negative and consequently the function $\varphi'_{x,p}(\lambda) = (f'(x + \lambda p), p)$ is nondecreasing.

Since it has been shown that (1) $\rightarrow$ (2) $\rightarrow$ (3) $\rightarrow$ (1), (4) $\rightarrow$ (3) and (3) $\rightarrow$ (4) the equivalence of all of the four statements in lemma 2.5 is proved.

Corollary 2.1. *The quadratic function*

$$f(x) = \frac{1}{2}(x, Ax) + (b, x)$$

is convex if and only if matrix A is positive definite.

Indeed, $f(x)$ is twice continuously differentiable and $f''(x) = A$. Therefore the statement of the corollary follows directly from statement (4) of lemma 2.5.

Lemma 2.5 provides a series of criteria of convexity of a function which enable us to establish whether a given function is convex.

Definition 2.2. *Let the convex function $f(x)$ be defined at point x_0 and have a finite value. Vector g is called a subgradient or support vector for function $f(x)$ at point x_0 if for any x the inequality*

$$f(x) - f(x_0) \geq (g, x - x_0) \quad (2.6)$$

is satisfied.

It can be shown that if $f(x)$ is continuous at point x_0 , then at this point there exist subgradients and the set of these subgradients is convex, closed and bounded. It follows from lemma 2.5 (statement 2)

that $f'(x_0)$ is a subgradient of function $f(x)$ at point x_0 if $f(x)$ is differentiable. Thus the concept of subgradient is a generalization of the gradient concept.

It is clear from the definition that if g_1 and g_2 are subgradients of convex functions $f_1(x)$ and $f_2(x)$ at point x_0 , then $c_1g_1 + c_2g_2$ is a subgradient of function $c_1f_1(x) + c_2f_2(x)$, $c_1, c_2 \geq 0$. Thus knowing the subgradients of certain convex functions it is easy to compute the subgradient for their linear combination.

Now let $f(x) = \max_{i=1, \dots, m} f_i(x)$, where $f_i(x)$ is a convex function, and let g_i be subgradients of $f_i(x)$ at point x_0 . Then vector

$$g = \sum_{i=1}^m \lambda_i g_i$$

where $\sum_{i=1}^m \lambda_i = 1$, $\lambda_i \geq 0$, $i = 1, \dots, m$, $\lambda_i = 0$ if $f_i(x_0) < f(x_0)$, is a subgradient of function $f(x)$.

Strictly and Strongly Convex Functions

Functions for which the condition of convexity is satisfied in a strong sense play a very important role in mathematical programming.

Definition 2.3. Function $f(x)$ is called strictly convex if

$$f((1 - \lambda)x + \lambda y) < (1 - \lambda)f(x) + \lambda f(y), \quad 0 < \lambda < 1, \\ x \neq y.$$

If a strictly convex function is sufficiently smooth, then statements similar to those formulated in lemma 2.5 are valid for it.

Lemma 2.6. The following statements are equivalent:

- (1) $f(x)$ is a strictly convex function.
- (2) $f(x_2) - f(x_1) > (f'(x_1), x_2 - x_1)$ for any $x_1, x_2 \in E^n$, $x_1 \neq x_2$.
- (3) $(f'(x + \lambda p), p)$ is a strictly increasing function of λ .

Definition 2.4. Function $f(x)$ is called strongly convex if for any $x_1, x_2 \in E^n$

$$f\left(\frac{x_1 + x_2}{2}\right) \leq \frac{1}{2}[f(x_1) + f(x_2)] - \gamma \|x_2 - x_1\|^2 \quad (2.7)$$

where $\gamma > 0$ is an arbitrary small constant.

A strongly convex function as can be easily ascertained is also strictly convex, but, speaking generally, the converse does not hold.

In what follows we shall consider twice continuously differentiable strongly convex functions.

Lemma 2.7. *If $f(x)$ is a twice continuously differentiable function, then the condition of strong convexity (2.7) is equivalent to the condition*

$$(f''(x)p, p) \geq m \|p\|^2, \quad m > 0 \quad (2.8)$$

for any x and $p \in E^n$.

Inequality (2.8) implies that matrix $f''(x)$ is strongly positive.

Corollary 2.2. *A strictly convex quadratic function $f(x) = \frac{1}{2}(Ax, x) + (b, x)$ defined over the space E^n is strongly convex too and the converse is valid.*

Proof. It is necessary to prove only the first statement.

From (2) of lemma 2.6 it follows that for any $x \neq 0$

$$(Ax, x) > 0. \quad (2.9)$$

At the same time

$$(Ax, x) \geq \lambda (x, x) = \lambda \|x\|^2 \quad (2.10)$$

where λ is the least eigenvalue of the matrix of second derivatives, A . From (2.9) and (2.10) it follows that $\lambda > 0$ and $f(x)$ is a strongly convex function.

Let x_0 be an arbitrary point in E^n . Consider the set

$$Y = \{x: f(x) \leq f(x_0)\}.$$

Lemma 2.8. *If $f(x)$ is a twice continuously differentiable strongly convex function, then Y is a closed bounded strongly convex set.*

Proof. The set Y is closed since $f(x)$ is a continuous function.

Let us prove that Y is bounded. By Taylor's formula

$$f(x) = f(x_0) + (f'(x_0), x - x_0) + \frac{1}{2}(f''(\xi)(x - x_0), x - x_0)$$

where $\xi = x_0 + \theta(x - x_0)$, $\theta \in [0, 1]$. Using (2.8) we have

$$f(x_0) \geq f(x) \geq f(x_0) + (f'(x_0), x - x_0) + \frac{m}{2} \|x - x_0\|^2.$$

Hence

$$\frac{m}{2} \|x - x_0\|^2 + (f'(x_0), x - x_0) < 0,$$

i.e.

$$\frac{m}{2} \|x - x_0\|^2 \leq |(f'(x_0), x - x_0)| \leq \|f'(x_0)\| \|x - x_0\|$$

or

$$\|x - x_0\| \leq 2 \frac{\|f'(x_0)\|}{m}.$$

This last inequality proves that Y is bounded.

Finally, let us establish that Y is a strongly convex set. Let $x_1, x_2 \in Y$. Using Lagrange's formula and condition (2.7) we obtain

$$\begin{aligned} f\left(\frac{x_1+x_2}{2}+y\right) &= f\left(\frac{x_1+x_2}{2}\right) + (f'(\xi), y) \\ &\leq \frac{1}{2}[f(x_1) + f(x_2)] - \gamma \|x_1 - x_2\|^2 + M \|y\| \end{aligned} \quad (2.11)$$

where $\xi = \frac{x_1+x_2}{2} + \theta y$, $\theta \in [0, 1]$, M is the maximum value of $f'(x)$, the derivative on the set Y .

Setting $f(x_1) \geq f(x_2)$ we have $\frac{1}{2}[f(x_1) + f(x_2)] \leq f(x_1)$. If $\|y\| \leq \frac{\gamma}{M} \|x_2 - x_1\|^2$, then from (2.11) it follows that $f\left(\frac{x_1+x_2}{2}+y\right) \leq f(x_1)$, i.e. $\frac{x_1+x_2}{2}+y \in Y$. By definition this means that Y is a strongly convex set.

The lemma is proved.

Remark. The set Y remains closed and strongly convex also in case $f(x)$ is a differentiable or continuous strongly convex function. The proof of Y being strongly convex is based on the fact that a continuous strongly convex function satisfies in every bounded set Lipschitz' condition (see N. Bourbaki).

Lemma 2.9. *If the matrix $f''(x)$ satisfies condition (2.8), then there exists the inverse matrix $f''^{-1}(x)$ and also*

$$(f''^{-1}(x) p, p) \leq \frac{1}{m} \|p\|^2.$$

If moreover the matrix $f''(x)$ is bounded, i.e.

$$(f''(x) p, p) \leq M \|p\|^2 \quad (2.12)$$

then

$$(f''^{-1}(x) p, p) \geq \frac{m}{M^2} \|p\|^2.$$

Concave Functions

Definition. *If for any $x_1, x_2 \in E^n$ and any $0 \leq \lambda \leq 1$ the inequality*

$$f(\lambda x_1 + (1 - \lambda) x_2) \geq \lambda f(x_1) + (1 - \lambda) f(x_2)$$

is satisfied, then the function $f(x)$ is called concave.

It follows that the function $f(x)$ is concave if and only if the function $-f(x)$ is convex. Taking this into account all the properties of concave function can be obtained by a simple reformulation of the corresponding properties of convex functions.

In a way analogous to that used for convex functions, strictly and strongly concave functions can be defined and their properties studied.

3. CONVEX PROGRAMMING

The subject matter of convex programming is minimization of a convex function in a convex domain. Convex programming is the most elaborated part of mathematical programming.

Formulation of the Problem. Basic Properties

Given a convex continuous function $f(x)$, $x \in E^n$, defined for all $x \in E^n$, and a convex set X . It is required to find the minimum of $f(x)$ in the set X , i.e. to find point x_* such that

$$f(x_*) \leq f(x), \quad x \in X.$$

Lemma 3.1. *A convex continuous function $f(x)$ attains its minimum in a compact convex set X .*

Proof. The hypothesis is just a particular case of the well-known Weierstrass theorem which states that a continuous function attains its minimum in a compact set.

Lemma 3.2. *Let X be a closed set and $f(x)$ a twice continuously differentiable strongly convex function. Then $f(x)$ attains its minimum in X .*

Proof. Let $x_0 \in X$. Consider the set

$$Y = \{x: f(x) \leq f(x_0)\}.$$

By lemma 2.8 it is closed and bounded. Consider now the intersection $X \cap Y$. Obviously, if x_* is the minimum of $f(x)$ in the set $X \cap Y$, then this point is the minimum point of $f(x)$ in X as well. But the set $X \cap Y$ is bounded and closed being the intersection of two closed sets one of which is bounded. Therefore $f(x)$ attains its minimum in $X \cap Y$ and consequently in X as a whole.

Convex and strictly convex functions can fail to attain their minimum.

Lemma 3.3. *A set of points $X_* \subset X$ at which the convex function $f(x)$ attains its minimum in X is convex.*

Lemma 3.4. *A strictly convex function attains its minimum in a convex set X at one and only one point.*

Proof. Let x_1 and x_2 be different points of minimum of $f(x)$ in X . Then

$$f\left(\frac{1}{2}x_1 + \frac{1}{2}x_2\right) < \frac{1}{2}f(x_1) + \frac{1}{2}f(x_2) = f(x_1),$$

$$\frac{1}{2}x_1 + \frac{1}{2}x_2 \in X.$$

This contradicts the fact that x_1 is a point of minimum of $f(x)$.

Necessary Conditions for a Minimum

Let $f(x)$ be a continuously differentiable convex function and X a convex set. We have to consider the following question: if x_* is the minimum point of $f(x)$ in X , what conditions are to be satisfied at this point?

Definition 3.1. Let $x_0 \in X$. We denote by $K(x_0)$ a set of vectors p such that $p \in K(x_0)$ if and only if there is an $\alpha > 0$ such that $x_0 + \alpha p \in X$.

The set $K(x_0)$ is called the cone of admissible directions for X at point x_0 .

Lemma 3.5. $K(x_0)$ is a convex cone. If $p \in K(x_0)$ and $x_0 + \alpha_0 p \in X$, then $x_0 + \alpha p \in X$ with any $0 \leq \alpha \leq \alpha_0$.

Theorem 3.1. Let x_* be the minimum point of a continuously differentiable convex function $f(x)$ in a convex set X . Then

$$f'(x_*) \in K^*(x_*). \quad (3.1)$$

Conversely if (3.1) holds, then x_* is the minimum point of $f(x)$ in X .

Proof. Let (3.1) be satisfied at point x_* . Then $(f'(x_*), p) \geq 0$, $p \in K(x_*)$. Further if $x \in X$, then $p = x - x_* \in K(x_*)$ for $x_* + (x - x_*) = x \in X$. Therefore

$$(f'(x_*), x - x_*) \geq 0, \quad x \in X.$$

By lemma 2.5 we have for a convex function

$$f(x) - f(x_*) \geq (f'(x_*), x - x_*).$$

Hence

$$f(x) - f(x_*) \geq 0, \quad x \in X$$

this shows that x_* is the minimum point of $f(x)$ in X .

Let us now prove that condition (3.1) is necessary. Let x_* be the minimum point. Then for any $x \in X$ and λ , $0 < \lambda \leq 1$, we have

$$f((1 - \lambda)x_* + \lambda x) = f(x_* + \lambda(x - x_*)) \geq f(x_*)$$

or

$$\frac{f(x_* + \lambda(x - x_*)) - f(x_*)}{\lambda} \geq 0.$$

In the limit with $\lambda \rightarrow 0$ we obtain

$$(f'(x_*), x - x_*) \geq 0, \quad x \in X. \quad (3.2)$$

Let now $p \in K(x_*)$. Then $x_* + \alpha p = x \in X$, $\alpha > 0$, or

$$p = \frac{1}{\alpha} (x - x_*).$$

Then by (3.2) and with $\alpha > 0$, we have

$$(f'(x_*), p) = \frac{1}{\alpha} (f'(x_*), x - x_*) \geq 0. \quad (3.3)$$

Inequality (3.3) is valid for any $p \in K(x_*)$. Hence

$$f'(x_*) \in K^*(x_*).$$

Corollary 3.1. *By theorem 3.1 point x_* is the minimum point of $f(x)$ in X if and only if the inequality*

$$(f'(x_*), x - x_*) \geq 0, \quad x \in X$$

is satisfied. In fact, as just shown, (3.2) is equivalent to (3.1).

Let us show how theorem 3.1 is applied to the case where domain X is defined by a system of linear inequalities.

Given vectors $a_i \in E^n$, $i \in \mathcal{J}^- \cup \mathcal{J}^0$, where $\mathcal{J}^-$ and $\mathcal{J}^0$ are finite sets of indices, and corresponding numbers b_i . Let domain X be defined by a system of equalities and inequalities:

$$(a_i, x) - b_i \leq 0, \quad i \in \mathcal{J}^-, \quad (a_i, x) - b_i = 0, \quad i \in \mathcal{J}^0. \quad (3.4)$$

Let us describe a cone $K(x_0)$ at an arbitrary point $x_0 \in X$. We set

$$\mathcal{J}^-(x_0) = \{i: (a_i, x_0) - b_i = 0, i \in \mathcal{J}^-\}.$$

By definition, $p \in K(x_0)$ if $x_0 + \alpha p \in X$ with sufficiently small α . It is clear that $x_0 + \alpha p \in X$, i.e. that point $x_0 + \alpha p$ satisfies (3.4) with small α if and only if

$$(a_i, p) \leq 0, \quad i \in \mathcal{J}^-(x_0), \quad (a_i, p) = 0, \quad i \in \mathcal{J}^0. \quad (3.5)$$

Thus cone $K(x_0)$ is described by system (3.5) which we can rewrite in an equivalent form:

$$(-a_i, p) \geq 0, \quad i \in \mathcal{J}^-(x_0), \quad (a_i, p) \geq 0, \quad i \in \mathcal{J}^0, \quad (-a_i, p) \geq 0, \quad i \in \mathcal{J}^0.$$

By lemma 1.6 vector $y \in K^*(x_0)$ can be presented in the form

$$y = \sum_{i \in \mathcal{J}^-(x_0)} -u^i a_i + \sum_{i \in \mathcal{J}^0} -u^{+i} a_i + \sum_{i \in \mathcal{J}^0} u^{-i} a_i,$$

where u^i , u^{+i} , u^{-i} are nonnegative numbers. Denoting $u^i = u^{+i} - u^{-i}$, $i \in \mathcal{J}^0$ we obtain

$$y = - \sum_{i \in \mathcal{J}^-(x_0)} u^i a_i - \sum_{i \in \mathcal{J}^0} u^i a_i, \quad u^i \geq 0, \quad i \in \mathcal{J}^-(x_0). \quad (3.6)$$

Theorem 3.2. *Let $f(x)$ be a convex differentiable function and set X be defined by system (3.4). Then for point x_* to be the minimum point of $f(x)$ in X it is necessary and sufficient that there exist numbers u^i , $i \in \mathcal{J}^- \cup \mathcal{J}^0$ such that*

$$f'(x_*) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i = 0, \quad u^i \geq 0, \quad i \in \mathcal{J}^-, \quad u^i = 0,$$

if $(a_i, x_*) - b_i < 0$, $i \in \mathcal{J}^-$.

Proof. The result is obtained directly by using theorem 3.1 and y in the form (3.6) for elements of $K^*(x_0)$ and also taking $u^i = 0$ for $i \in \mathcal{J}^-(x_*)$.

Corollary 3.2. *For point x_* to be the minimum point of the convex differentiable function over the whole space it is necessary and sufficient to satisfy the equality*

$$f'(x_*) = 0.$$

Corollary 3.3. *For point x_* to be the minimum point of the convex differentiable function in the set*

$$x^j \geq 0, \quad j \in \mathcal{Y},$$

where $\mathcal{Y}$ is a subset of the set $j = 1, 2, \dots, n$, it is necessary and sufficient to satisfy the relations

$$\begin{aligned} \frac{\partial f(x_*)}{\partial x^j} &\geq 0 \quad \text{if } x^j = 0, \quad j \in \mathcal{Y}, \\ \frac{\partial f(x_*)}{\partial x^j} &= 0 \quad \text{if } x^j \neq 0 \text{ or } j \notin \mathcal{Y}. \end{aligned}$$

The Kuhn-Tucker Theorem

The necessary and sufficient conditions for a minimum considered above were based on an abstract description of an admissible set X in which function $f(x)$ was minimized. In a broad class of problems the set is defined by a system of inequalities and equalities. This section considers the necessary conditions for a minimum in this concrete case.

Given convex functions $f_i(x)$, $i = 0, 1, \dots, m$ and convex set X . It is required to minimize $f_0(x)$ with the following constraints

$$f_i(x) \leq 0, \quad i = 1, \dots, m, \quad x \in X. \quad (3.7)$$

Theorem 3.3 (Kuhn-Tucker). *Let x_* be the minimum point of $f_0(x)$ with the constraints (3.7) and let there be a point $x_1 \in X$ such that*

$$f_i(x_1) < 0, \quad i = 1, \dots, m.$$

Then there are numbers $u^i \geq 0$, $i = 1, \dots, m$ such that

$$\begin{aligned} f_0(x_*) + \sum_{i=1}^m u^i f_i(x_*) &\leq f_0(x) + \sum_{i=1}^m u^i f_i(x), \quad x \in X, \\ u^i f_i(x_*) &= 0, \quad i = 1, \dots, m. \end{aligned} \quad (3.8)$$

These conditions are necessary and sufficient.

Definition 3.2. Numbers u^i used in the theorem are called Lagrange multipliers.

Dual Problem

Consider again the problem of minimization of convex function $f_0(x)$ with constraints (3.7). Let $u^i \geq 0$, $i = 1, \dots, m$ be fixed. Let us compute

$$\varphi(u) = \inf_{x \in X} [f_0(x) + \sum_{i=1}^m u^i f_i(x)]. \quad (3.9)$$

Thus function $\varphi(u)$ with $u \geq 0$ has been determined; it can take the value $-\infty$ as well. We leave it to the reader to prove that $\varphi(u)$ is a concave function.

Theorem 3.4. Let $u \geq 0$ and x satisfy the constraints (3.7). Then

$$\varphi(u) \leq f_0(x).$$

If however the conditions of theorem 3.3 are satisfied, then

$$\max_{u \geq 0} \varphi(u) = \min_{x \in D} f_0(x)$$

where D is a set of points x which satisfy (3.7).

Proof. For $x \in D$, $u \geq 0$ we have

$$\varphi(u) \leq f_0(x) + \sum_{i=1}^m u^i f_i(x) \leq f_0(x).$$

Let now the conditions of theorem 3.3 be satisfied. Then there exists a vector $u_0 \geq 0$ such that the relations (3.8) are satisfied for it. These relations imply

$$\varphi(u_0) = f_0(x_*) + \sum_{i=1}^m u_0^i f_i(x_*) = f_0(x_*),$$

and since $\varphi(u) \leq f_0(x_*)$ it follows that vector u_0 provides the maximum of function $\varphi(u)$ in the domain $u \geq 0$ and

$$\max_{u \geq 0} \varphi(u) = \varphi(u_0) = f_0(x_*) = \min_{x \in D} f_0(x).$$

Q.E.D.

The problem of maximization of $\varphi(u)$ with the constraint $u \geq 0$ is known as the dual problem of convex programming and u as the vector of dual variables.

The essence of theorem 3.4 can now be interpreted as follows: under the conditions of the Kuhn-Tucker theorem the value of the maximum of the objective function in the dual problem is that of the minimum of the objective function of the primal problem. The Lagrange multipliers of the primal problem are at the same time the solution of the dual problem.

The problem of convex programming often arises in the form of the minimization of $f_0(x)$ with the constraints

$$f_i(x) \leq 0, \quad i \in \mathcal{J}^-, \quad f_i(x) = 0, \quad i \in \mathcal{J}^0, \quad x \in X \quad (3.10)$$

where $\mathcal{J}^-$ and $\mathcal{J}^0$ are finite sets of indices, $f_0(x)$ and $f_i(x)$, $i \in \mathcal{J}^-$, are convex functions of x , $f_i(x)$, $i \in \mathcal{J}^0$ are linear functions and X is a convex set.

The dual problem for this case is formulated as a maximization problem of $\varphi(u)$ with the constraints $u^i \geq 0$, $i \in \mathcal{J}^-$ where u has components u^i , $i \in \mathcal{J}^- \cup \mathcal{J}^0$ and

$$\varphi(u) = \inf_{x \in X} [f_0(x) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i f_i(x)]. \quad (3.11)$$

Thus the number of dual variables is equal to the number of constraints (3.10) and the variable u^i corresponding to the i -th constraint takes nonnegative values if it corresponds to an inequality constraint and arbitrary values if it corresponds to an equality constraint.

Problem of Linear Programming

The problem of linear programming is the problem of minimization of the function $f_0(x) = (a_0, x)$ subject to constraints (3.4)

$$(a_i, x) - b_i \leq 0, \quad i \in \mathcal{J}^-, \quad (a_i, x) - b_i = 0, \quad i \in \mathcal{J}^0.$$

This problem coincides with the problem (3.10) if

$$f_i(x) = (a_i, x) - b_i, \quad X \in E^n.$$

Lemma 3.6. *If constraints (3.4) apply together, then the problem of linear programming either has a solution x_* or the value of the lower bound $f_0(x) = (a_0, x)$ with the constraints (3.4) is $-\infty$.*

The proof of this lemma can be found in textbooks on linear programming.

The necessary conditions characterizing x_* , the solution of the problem of linear programming, are obtained just by reformulating theorem 3.2 since $f'_0(x) = a_0$.

Theorem 3.5. *In order that point x_* be the solution of a problem of linear programming it is necessary and sufficient that there be numbers u^i , $i \in \mathcal{J}^- \cup \mathcal{J}^0$ such that*

$$a_0 + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i = 0, \quad u^i \geq 0, \quad i \in \mathcal{J}^-, \quad u^i = 0, \quad (3.12)$$

if $(a_i, x_*) - b_i < 0$, $i \in \mathcal{J}^-$.

Let us construct the dual of the problem of linear programming. By definition,

$$\begin{aligned} \varphi(u) &= \inf_{x \in E^n} [f_0(x) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i f_i(x)] \\ &= \inf_{x \in E^n} [(a_0, x) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i ((a_i, x) - b_i)] \\ &= \inf_{x \in E^n} [((a_0 + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i), x) - \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i b_i] \\ &= \begin{cases} - \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i b_i & \text{if } a_0 + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i = 0, \\ -\infty & \text{if } a_0 + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i \neq 0. \end{cases} \end{aligned}$$

Thus the dual of the problem of linear programming, i.e. the problem of the maximization of $\varphi(u)$ with $u^i \geq 0$, $i \in \mathcal{J}^-$, is equivalent to the maximization of

$$- \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i b_i \quad (3.13)$$

with the constraints

$$a_0 + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i = 0, \quad u^i \geq 0, \quad i \in \mathcal{J}^-. \quad (3.14)$$

Theorem 3.6. *If the primal problem of linear programming has a solution, then the Lagrange multipliers are the solution of the dual problem, and at the same time the value of the minimum of the objective function of the primal problem is equal to the value of the maximum of the objective function of the dual problem.*

In addition to constraints (3.4), problems of linear programming often contain constraints of the type

$$x^j \geq 0, \quad j \in \mathcal{Y} \quad (3.4')$$

where $\mathcal{Y}$ is a subset of the set $j = 1, 2, \dots, n$. Using theorem 3.6 the reader can easily prove the following.

Theorem 3.7. *If a problem of linear programming with constraints (3.4), (3.4') has a solution, then the Lagrange multipliers corresponding*

to constraints (3.4) are the solution of the dual problem: the maximization of

$$-\sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i b_i$$

with constraints

$$\begin{aligned} a_0^j + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i^j &\geq 0, \quad j \in \mathcal{Y}, \\ a_0^j + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i^j &= 0, \quad j \notin \mathcal{Y}, \quad u^i \geq 0, \quad i \in \mathcal{J}^-, \end{aligned}$$

where a_i^j is the j -th component of vector a_i . The values of the minimum of the objective function of the primal problem and of the maximum of the dual one coincide.

Problem of Quadratic Programming

The problem of quadratic programming consists in the minimization of the quadratic function

$$f_0(x) = \frac{1}{2}(x, Cx) + (d, x)$$

with constraints (3.4); here C is an $n \times n$ symmetric, positive definite matrix, and d is an n -dimensional vector.

Lemma 3.7. *In a problem of quadratic programming the lower bound is either attained or is $-\infty$.*

The proof of this lemma will be omitted.

Theorem 3.8. *In order that point x_* be the solution of a problem of quadratic programming, it is necessary and sufficient that there be numbers u_*^i , $i \in \mathcal{J}^- \cup \mathcal{J}^0$ such that*

$$Cx_* + d + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u_*^i a_i = 0,$$

$$u_*^i = 0 \quad \text{if} \quad (a_i, x_*) - b_i < 0, \quad i \in \mathcal{J}^-; \quad u_*^i \geq 0, \quad i \in \mathcal{J}^-.$$

The hypothesis can be proved by directly using theorem 3.2.

Let now matrix C be strictly positive definite, i.e. there is a γ such that $(x, Cx) \geq \gamma \|x\|^2$. In this case matrix C is nonsingular and has an inverse matrix C^{-1} . Let us construct the dual problem:

$$\begin{aligned} \varphi(u) &= \inf_{x \in E^n} [f_0(x) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i f_i(x)] \\ &= \inf_{x \in E^n} \left[\frac{1}{2}(x, Cx) + (d, x) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i ((a_i, x) - b_i) \right] \\ &= \inf_{x \in E^n} \left[- \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i b_i + \frac{1}{2}(x, Cx) + \left(x, d + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i \right) \right]. \end{aligned}$$

Equating the derivatives of the right-hand side to zero, we find that the minimum is attained with

$$x(u) = -C^{-1} \left(d + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i \right).$$

At the same time

$$\begin{aligned} \varphi(u) = & - \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i b_i \\ & - \frac{1}{2} \left(d + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i, C^{-1} \left(d + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i \right) \right). \end{aligned} \quad (3.15)$$

Thus the dual problem consists in the maximization of (3.15) with constraints $u^i \geq 0$, $i \in \mathcal{J}^-$.

Theorem 3.9. *If the minimum in the problem of quadratic programming is attained and matrix C is strictly positive definite, then the Kuhn-Tucker theorem and theorem 3.4 are valid for the problem of quadratic programming. In this case the Lagrange multipliers of the primal problem are the solution of the dual problem, and if u_* is the solution of the dual problem, then the solution of the primal one can be found by the following formula:*

$$x(u) = -C^{-1} \left(d + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i \right). \quad (3.16)$$

4. NECESSARY CONDITIONS FOR A MINIMUM

The general problem of mathematical programming consists in minimizing function $f_0(x)$, $x \in E^n$ in a set defined by a system of equalities and inequalities

$$f_i(x) \leq 0, \quad i \in \mathcal{J}^-, \quad f_i(x) = 0, \quad i \in \mathcal{J}^0, \quad x \in X, \quad (4.1)$$

where $\mathcal{J}^-$ and $\mathcal{J}^0$ are finite sets of indices. In this section it is always assumed that $f_i(x)$ are continuously differentiable functions whose gradient is $f'_i(x)$. No assumption is made about set X for the present.

The main object of this section is to deduce the necessary conditions which must be satisfied at point x_* providing the minimum of $f_0(x)$ subject to constraints (4.1).

Basic Definitions

Definition 4.1. *A set D of points which satisfy constraints (4.1) is called an admissible domain.*

We assume that this set is nonempty.

Definition 4.2. Function $f_0(x)$ being minimized in D is called an objective function.

Definition 4.3. Point x_* satisfying (4.1) for which

$$f_0(x_*) \leq f_0(x), \quad x \in D,$$

is called the minimum point.

Definition 4.4. Point x_* is called a point of local minimum of $f_0(x)$ in D if there is a neighbourhood Ω of point x_* such that

$$f(x_*) \leq f_0(x), \quad x \in D \cap \Omega.$$

In what follows the problem of the minimization of $f_0(x)$ will generally be considered. Obviously, the problem of the maximization of a function $f(x)$ in D is reduced to that of the minimization in D of the function $f_0(x) = -f(x)$.

Necessary Conditions for a Minimum

Definition 4.5. Vector $p \in E^n$ defines the admissible direction with respect to set X at point $x_0 \in X$ if for any vector $e_i \in E^n$, $i \in \mathcal{J}^0$ and any function $r^i(\lambda)$, $i \in \mathcal{J}^0$ which satisfies the condition

$$\lim_{\lambda \rightarrow 0} \frac{r^i(\lambda)}{\lambda} = 0 \quad (4.2)$$

the expression

$$x_0 + \lambda p + \sum_{i \in \mathcal{J}^0} r^i(\lambda) e_i \in X \quad (4.3)$$

is valid with sufficiently small $\lambda > 0$.

The basic result to be proved in this section can now be formulated.

Theorem 4.1. Let x_* be a point of local minimum of $f_0(x)$ in D . Besides, let the set of admissible directions with respect to set X at point x_* form a convex cone $K(x_*)$. Then there are numbers $u^0, u^i \in \mathcal{J}^- \cup \mathcal{J}^0$ such that

$$\begin{aligned} u^0 f'_0(x_*) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i f'_i(x_*) &\in K^*(x_*), \\ u^i f_i(x_*) &= 0, \quad i \in \mathcal{J}^-, \quad u^i \geq 0, \quad i = 0, \quad i \in \mathcal{J}^-. \end{aligned} \quad (4.4)$$

Proof. Consider two cases.

(1) Vectors $f'_i(x_*)$, $i \in \mathcal{J}^0$ are linearly dependent. Then there are numbers u^i , $i \in \mathcal{J}^0$ such that

$$\sum_{i \in \mathcal{J}^0} u^i f'_i(x_*) = 0.$$

Taking $u^0 = 0$, $u^i = 0$, $i \in \mathcal{J}^-$ we see that all the conditions of theorem (4.1) are satisfied.

(2) Vectors $f'_i(x_*)$, $i \in \mathcal{J}^0$ are linearly independent. Then there are vectors e_i , $i \in \mathcal{J}^0$ such that

$$(f'_i(x_*), e_j) = \delta_{ij}, \quad i, j \in \mathcal{J}^0$$

where $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ii} = 1$.

Let the total number of indices i in the set $\mathcal{J}^- \cup \mathcal{J}^0$ be m . Consider set Z in space E^{m+1} , the set being defined as follows. Vector z belongs to Z if and only if there is a vector $p \in K(x_*)$ such that $z^i = (f'_i(x_*), p)$ if $f_i(x_*) = 0$, $i \in \mathcal{J}^- \cup \mathcal{J}^0$ or $i = 0$. The components z^i (of vector $z \in Z$) for which $f_i(x_*) < 0$ are arbitrary. Since $K(x_*)$ is a convex cone it is easily seen that Z is a convex cone too.

Let us now define the set P . Vector w belongs to P if and only if

$$\begin{aligned} w^i &< 0 \text{ with } f_i(x_*) = 0, \quad i \in \mathcal{J}^- \text{ or } i = 0 \\ w^i &= 0 \text{ with } i \in \mathcal{J}^0. \end{aligned}$$

The remaining components of vector w are arbitrary. Obviously, P is a convex set too.

Let us demonstrate that Z and P do not intersect. Suppose that the opposite is true. Then there must be a vector $p_0 \in K(x_*)$ such that

$$(f'_i(x_*), p_0) < 0 \text{ if } i \in \mathcal{J}^- \text{ and } f_i(x_*) = 0$$

and

$$(f'_i(x_*), p_0) = 0 \text{ with } i \in \mathcal{J}^0. \quad (4.5)$$

We now construct a system of equations in functions $r^i(\lambda)$, $i \in \mathcal{J}^0$:

$$f_i(x_* + \lambda p_0 + \sum_{i \in \mathcal{J}^0} r^i e_i) = 0, \quad i \in \mathcal{J}^0. \quad (4.6)$$

Let us denote

$$g_i(\lambda, r) \equiv f_i(x_* + \lambda p_0 + \sum_{i \in \mathcal{J}^0} r^i e_i), \quad i \in \mathcal{J}^0.$$

Then from (4.6) we have

$$g_i(\lambda, r) = 0, \quad i \in \mathcal{J}^0 \quad (4.7)$$

which defines r^i as implicit functions of λ . Since it was assumed that $f_i(x)$ are continuously differentiable functions, the functions $g_i(\lambda, r)$ are also continuously differentiable in λ and r^i . Then using (4.5) we write

$$\frac{\partial g_i(0, 0)}{\partial \lambda} = (f'_i(x_*), p_0) = 0, \quad i \in \mathcal{J}^0, \quad (4.8)$$

$$\frac{\partial g_i(0, 0)}{\partial r^j} = (f'_i(x_*), e_j) = \delta_{ij}. \quad (4.9)$$

Let us denote by $\frac{\partial g}{\partial r}$ a matrix whose components are $\frac{\partial g_i(0, 0)}{\partial r^j}$, $i, j \in \mathcal{J}^0$. By the theorem on implicit functions, system (4.7) can be solved for r with small λ if matrix $\frac{\partial g}{\partial r}$ is nonsingular. In this case $r(\lambda)$ is a differentiable function of λ , $r(0) = 0$ and

$$r'(\lambda) = - \left(\frac{\partial g}{\partial r} \right)^{-1} \frac{\partial g}{\partial \lambda} \quad (4.10)$$

where $\frac{\partial g}{\partial \lambda}$ is a vector whose components are $\frac{\partial g_i(0, 0)}{\partial \lambda}$. In the case under consideration

$$\frac{\partial g}{\partial r} = I, \quad \frac{\partial g}{\partial \lambda} = 0, \quad (4.11)$$

where I is an identity matrix. This follows from (4.8) and (4.9).

Thus we see that with small λ continuously differentiable functions $r^i(\lambda)$, $i \in \mathcal{J}^0$ are defined. From (4.10) and (4.11) we have

$$\lim_{\lambda \rightarrow 0} \frac{r^i(\lambda) - r^i(0)}{\lambda} = \lim_{\lambda \rightarrow 0} \frac{r^i(\lambda)}{\lambda} = r^{i'}(0) = 0 \quad (4.12)$$

Let now $x(\lambda) = x_* + \lambda p_0 + \sum_{i \in \mathcal{J}^0} r^i(\lambda) e_i$. Then $x(\lambda) \in X$ with small $\lambda > 0$ by the definition of $K(x_*)$, for $p_0 \in K(x_*)$. Further, $f_i(x(\lambda)) = 0$, $i \in \mathcal{J}^0$, since $r^i(\lambda)$ satisfy (4.6) by definition. Further $f_0(x(\lambda)) < f_0(x_*)$ with small $\lambda > 0$. In fact by Taylor's formula

$$f_0(x(\lambda)) = f_0(x_*) + (f'_0(\xi), x(\lambda) - x_*)$$

where ξ is a point of the segment joining x_* and $x(\lambda)$. Therefore

$$\frac{f_0(x(\lambda)) - f_0(x_*)}{\lambda} = (f'_0(\xi), p_0) + \sum_{i \in \mathcal{J}^0} \frac{r^i(\lambda)}{\lambda} (f'_0(\xi), e_i).$$

Since from (4.5) $(f'_0(x_*), p_0) < 0$ and $\frac{r^i(\lambda)}{\lambda} \rightarrow 0$ we obtain, with small positive λ , $\xi \rightarrow x_*$ and

$$\frac{f_0(x(\lambda)) - f_0(x_*)}{\lambda} < 0.$$

Similarly if $i \in \mathcal{J}^-$ and $f_i(x_*) = 0$, then by (4.5)

$$f_i(x(\lambda)) < 0, \quad i \in \mathcal{J}^-, \quad f_i(x_*) = 0.$$

If $f_i(x_*) < 0$, $i \in \mathcal{J}^-$, then $f_i(x(\lambda)) < 0$, by continuity.

Thus point $x(\lambda)$ with small positive λ satisfies all constraints (4.1) and $f_0(x(\lambda)) < f_0(x_*)$. But this contradicts the fact that x_* is a point of local minimum.

The contradiction obtained shows that sets Z and P do not intersect. Since these sets are convex, they can be separated. This means that there are numbers $u^0, u^i \in \mathcal{J}^- \cup \mathcal{J}^0$, not all of them zero, such that

$$u^0 z^0 + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i z^i \geq u^0 w^0 + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i w^i, \quad z \in Z, w \in P. \quad (4.13)$$

The structure of sets Z and P makes it possible to draw certain conclusions about numbers u^i . By the definition of P , w^0 can take any value less than zero. Hence $u^0 \geq 0$, otherwise the right-hand side could take any great value and this contradicts (4.13). Similarly,

$$u^i \geq 0 \text{ if } f_i(x_*) = 0, \quad i \in \mathcal{J}^-. \quad (4.14)$$

Further, if $i \in \mathcal{J}^-$ and $f_i(x_*) < 0$, then w^i is arbitrary. Therefore for the inequality (4.13) to be valid it is necessary that

$$u^i = 0 \text{ if } f_i(x_*) < 0, \quad i \in \mathcal{J}^-. \quad (4.15)$$

Letting now w in (4.13) tend to zero so that $w \in P$ and taking into account (4.15) and the definition of Z , we obtain

$$u^0 (f'_0(x_*), p) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i (f'_i(x_*), p) \geq 0, \quad p \in K(x_*)$$

or

$$(u^0 f'_0(x_*) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i f'_i(x_*), p) \geq 0, \quad p \in K(x_*). \quad (4.16)$$

The statements proved (4.14), (4.15), (4.16) are obviously equivalent to the above theorem. This completes the proof.

Corollary 4.1. *If $X = E^n$, then for point x_* to provide a local minimum it is necessary that there be numbers u^i , not all zero, such that*

$$\begin{aligned} u^0 f'_0(x_*) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i f'_i(x_*) &= 0, \\ u^0 \geq 0, \quad u^i \geq 0, \quad i \in \mathcal{J}^-, \quad u^i f_i(x_*) &= 0, \quad i \in \mathcal{J}^-. \end{aligned} \quad (4.17)$$

Proof. If $X = E^n$, then any direction p is admissible, i.e. $K(x_*) = E^n$. Therefore cone $K^*(x_*)$ consists of one and only one zero vector and relations (4.4) directly take the form (4.17).

Corollary 4.2. For point x_* to provide the minimum of $f_0(x)$ in the domain

$$x^j \geq 0, \quad j \in \mathcal{Y}$$

where $\mathcal{Y}$ is a subset of the set $j = 1, 2, \dots, n$ it is necessary to satisfy the conditions

$$\begin{aligned} \frac{\partial f_0(x_*)}{\partial x^j} &\geq 0 \quad \text{if } x_*^j = 0, \quad j \in \mathcal{Y}, \\ \frac{\partial f_0(x_*)}{\partial x^j} &= 0 \quad \text{if } x_*^j > 0, \quad j \in \mathcal{Y} \text{ or } j \notin \mathcal{Y}. \end{aligned} \quad (4.18)$$

Proof. Constraints $x^j \geq 0, j \in \mathcal{Y}$ can be rewritten in the form $(-a_j, x) \leq 0, j \in \mathcal{Y}$, where a_j is a vector whose components are $a_j^i = \delta_{ij}, i = 1, \dots, n$. Using the preceding corollary we obtain that there are numbers u^0 and $u^j, j \in \mathcal{Y}$, not all zero such that

$$u^0 f'_0(x_*) - \sum_{j \in \mathcal{Y}} u^j a_j = 0, \quad u^0, \quad u^j \geq 0, \quad u^j x_*^j = 0, \quad j \in \mathcal{Y}. \quad (4.19)$$

The first equality can be written in terms of components in the following form:

$$u^0 \frac{\partial f_0(x_*)}{\partial x^i} - \sum_{j \in \mathcal{Y}} u^j \delta_{ij} = 0$$

or

$$u^0 \frac{\partial f_0(x_*)}{\partial x^i} = u^i, \quad i \in \mathcal{Y}, \quad u^0 \frac{\partial f_0(x_*)}{\partial x^i} = 0, \quad i \notin \mathcal{Y}. \quad (4.20)$$

It follows from (4.20) that $u^0 > 0$, since if $u^0 = 0$ then all $u^i = 0$, but this contradicts corollary (4.1). Therefore we can assume that $u^0 = 1$.

It follows directly from (4.20) and (4.19) that corollary 4.2 is valid.

Definition 4.6. The point x_* providing the minimum of $f_0(x)$ with constraints (4.1), where $X = E^n$, is called a regular point if gradients $f'_i(x_*)$ for indices i such that $i \in \mathcal{J}^- \cup \mathcal{J}^0, f_i(x_*) = 0$ are linearly independent.

Corollary 4.3. If x_* is a regular point, then in (4.17) we can take $u^0 = 1$ and the multipliers $u^i, i \in \mathcal{J}^- \cup \mathcal{J}^0$ are unique.

Proof. In fact, $u^0 > 0$. Since if $u^0 = 0$ then by (4.17) the gradients $f'_i(x_*)$ for which $i \in \mathcal{J}^- \cup \mathcal{J}^0, f_i(x_*) = 0$ would prove linearly dependent. Further, by (4.17) $u^i = 0$ if $f_i(x_*) < 0$. Therefore the

first of relations (4.17) with $u^0 = 1$ yields the expansion

$$f'_0(x_*) = - \sum_{\substack{i \in \mathcal{J}^- \cup \mathcal{J}^0 \\ f_i(x_*) = 0}} u^i f'_i(x_*)$$

of vector $f'_0(x_*)$ in linearly independent vectors $f'_i(x_*)$ and defines uniquely u^i .

Let there be now only equality constraints in problem (4.1)

$$f_i(x) = 0, \quad i \in \mathcal{J}^0$$

and $X = E^n$. If with such constraints x_* provides the minimum of $f_0(x)$ and gradients $f'_i(x_*)$ are linearly independent, then the necessary conditions for a minimum (4.17) can be written in the form

$$f'_0(x_*) + \sum_{i \in \mathcal{J}^0} u^i f'_i(x_*) = 0.$$

The set of vectors p such that

$$(f'_i(x_*), p) = 0, \quad i \in \mathcal{J}^0$$

in the case under consideration is called a tangent (bounding) manifold at point x_* to the set

$$D = \{x: f_i(x) = 0, \quad i \in \mathcal{J}^0\}.$$

Corollary 4.4. *For point x_* at which $f'_i(x_*)$, $i \in \mathcal{J}^0$ are linearly independent to provide the minimum of the function $f_0(x)$ in set D it is necessary that gradient $f'_0(x_*)$ be orthogonal to a manifold tangent to D at point x_* , i.e. if p belongs to the bounding manifold then $(f'_0(x_*), p) = 0$. In other words the projection of vector $f'_0(x_*)$ on the tangent manifold vanishes.*

Proof. If x_* with the above assumptions provides the minimum then

$$(f'_0(x_*), p) = - \sum_{i \in \mathcal{J}^0} u^i (f'_i(x_*), p) = 0$$

for any vector p of the tangent manifold. Conversely if $(f'_0(x_*), p)$ is equal to zero with any p , which belongs to the bounding manifold, then we can write

$$f'_0(x_*) = - \sum_{i \in \mathcal{J}^0} u^i f'_i(x_*)$$

This follows from lemma 1.6 if each of the equalities $(f'_i(x_*), p) = 0$ is written down in the form of two inequalities:

$$(f'_i(x_*), p) \geq 0, \quad -(f'_i(x_*), p) \geq 0.$$

Minimax Problem

It is required to find the minimum point of the function

$$f(x) = \max_{i=1, \dots, m} f_i(x) \quad (4.21)$$

where $f_i(x)$ are continuously differentiable functions, $x \in E^n$. In order to apply the results obtained in the preceding subsection let us reduce the problem of minimization of $f(x)$ to the equivalent problem of mathematical programming. It is easily seen that if we introduce a supplementary variable x^{n+1} , then x_* —the point of minimum of $f(x)$ —will also be the solution of the following problem: to find the minimum of $g_0(x, x^{n+1}) = x^{n+1}$ with the constraints

$$g_i(x, x^{n+1}) \equiv f_i(x) - x^{n+1} \leq 0, \quad i = 1, \dots, m. \quad (4.22)$$

The minimum value of $g_0(x, x^{n+1})$ is $x_*^{n+1} = f(x_*)$.

Let us apply corollary 4.1 of theorem 4.1 to problem (4.22). It is necessary to take into account that the problem will now be solved in space E^{n+1} of variables $x^1, \dots, x^n, x^{n+1}$ so that the gradients of the functions $g_i(x, x^{n+1})$ have the form

$$g'_i(x, x^{n+1}) = \begin{pmatrix} f'_i(x) \\ -1 \end{pmatrix}, \quad i = 1, \dots, m, \quad g'_0(x, x^{n+1}) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

By corollary 4.1 we now have: there are numbers $u^0, u^i, i = 1, \dots, m$ not all zero such that

$$\begin{aligned} u^0 \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \sum_{i=1}^m u^i \begin{pmatrix} f'_i(x_*) \\ -1 \end{pmatrix} &= 0, \\ u^i &\geq 0, \quad i = 0, 1, \dots, m, \\ u^i (f_i(x_*) - x_*^{n+1}) &= u^i (f_i(x_*) - f(x_*)) = 0, \quad i = 1, \dots, m. \end{aligned} \quad (4.23)$$

The first of relations (4.22) shows that $u^0 = \sum_{i=1}^m u^i$. Hence, since $u^i \geq 0$, we have $u^0 > 0$ for with $u^0 = 0$ all u^i would also be zero. Since expression (4.22) is homogeneous with respect to u^i we can take $u^0 = 1$.

Thus we have finally obtained the following result.

Theorem 4.2. *For point x_* to provide the minimum of $f(x)$ defined by relation (4.21) it is necessary that there be numbers $u^i, i = 1, \dots, m$ such that*

$$\begin{aligned} \sum_{i=1}^m u^i f'_i(x_*) &= 0, \\ \sum_{i=1}^m u^i &= 1, \quad u^i \geq 0, \quad i = 1, \dots, m, \\ u^i (f_i(x_*) - f(x_*)) &= 0, \quad i = 1, \dots, m. \end{aligned} \quad (4.24)$$

Necessary Conditions of the Second Order

Let us again return to the problem of the minimization of $f_0(x)$ with constraints (4.1), $X = E^n$. We use the notation $L(x, u)$ as follows:

$$L(x, u) = f_0(x) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i f_i(x). \quad (4.25)$$

Assume that point x_* —the solution of this minimization problem—is regular (definition 4.6). Then using corollary 4.3 of theorem 4.1 we can write the first of relations (4.17) in the form

$$L'_x(x_*, u) = 0. \quad (4.26)$$

Assume now that all functions $f_i(x)$ are twice continuously differentiable, i.e. that there are continuous matrices of second derivatives $f''_i(x)$. Then the matrix of second derivatives $L''_{xx}(x_*, u)$ of function $L(x, u)$ with respect to x is also defined.

The assumption that x_* is a regular point implies that the relation (4.17) is uniquely determined by multipliers u^i , $i \in \mathcal{J}^- \cup \mathcal{J}^0$. We introduce the following notations

$$\begin{aligned} \mathcal{J}_0^-(x_*) &= \{i: u^i > 0, i \in \mathcal{J}^-\}, \\ \mathcal{J}^-(x_*) &= \{i: f_i(x_*) = 0, i \in \mathcal{J}^-\}. \end{aligned}$$

Using (4.17) we have $\mathcal{J}_0^-(x_*) \subset \mathcal{J}^-(x_*)$. Let vector p satisfy the relations

$$\begin{aligned} (f'_i(x_*), p) &\leq 0, \quad i \in \mathcal{J}^-(x_*), \quad i \notin \mathcal{J}_0^-(x_*), \\ (f'_i(x_*), p) &= 0, \quad i \in \mathcal{J}_0^-(x_*) \cup \mathcal{J}^0. \end{aligned} \quad (4.27)$$

Assume

$$\mathcal{J}_p(x_*) = \{i \in \mathcal{J}^-(x_*) \cup \mathcal{J}^0: (f'_i(x_*), p) = 0\}. \quad (4.28)$$

Point x_* being regular, vectors $f'_i(x_*)$, $i \in \mathcal{J}_p(x_*)$ are linearly independent. Therefore it can be demonstrated that there is a function $r(\lambda) \in E^n$ such that

$$f_i(x(\lambda)) = 0, \quad i \in \mathcal{J}_p(x_*) \quad (4.29)$$

where $x(\lambda) = x_* + \lambda p + r(\lambda)$, $\lim_{\lambda \rightarrow 0} \frac{r(\lambda)}{\lambda} = 0$. The proof is quite similar to that of theorem 4.1.

Further, if $i \in \mathcal{J}_p(x_*)$, then either $f_i(x_*) < 0$ or $(f'_i(x_*), p) < 0$ which ensures the inequality $f_i(x(\lambda)) < 0$ with small λ . Thus point $x(\lambda)$ with small λ satisfies all the constraints (4.1), $X = E^n$. Using this fact as well as (4.27)-(4.29) we obtain

$$f_0(x(\lambda)) = L(x(\lambda), u)$$

since if $u^i \neq 0$ then from (4.29) $f_i(x(\lambda)) = 0$. At the same time we obtain from (4.17) that $f_0(x_*) = L(x_*, u)$. Taking into account that $x(\lambda)$ satisfies all of the relations (4.1) and that x_* is the minimum point of $f_0(x)$ with constraints (4.1) we obtain for small λ :

$$L(x(\lambda), u) \geq L(x_*, u).$$

Expanding $L(x(\lambda), u)$ to second-order terms in powers of λ we obtain

$$\begin{aligned} L(x(\lambda), u) &= L(x_*, u) + (L'_x(x_*, u), x(\lambda) - x_*) \\ &+ \frac{1}{2} (L''_{xx}(\xi(\lambda), u) (x(\lambda) - x_*), x(\lambda) - x_*) \geq L(x_*, u) \end{aligned}$$

where $\xi(\lambda)$ is a point in the segment which joins x_0 and $x(\lambda)$ so that $\xi(\lambda) \rightarrow x_*$ as $\lambda \rightarrow 0$. Using (4.26) we obtain

$$\frac{1}{2} \lambda^2 \left(L''_{xx}(\xi(\lambda), u) \left(p + \frac{r(\lambda)}{\lambda} \right), p + \frac{r(\lambda)}{\lambda} \right) \geq 0.$$

Dividing by λ^2 and taking $\lambda \rightarrow 0$ we finally obtain

$$(L''_{xx}(x_*, u) p, p) \geq 0.$$

The following theorem is proved.

Theorem 4.3. *Let functions $f_i(x)$ be twice continuously differentiable and x_* be a regular point of minimum of $f_0(x)$ with constraints (4.1), $X = E^n$. Then there are numbers u^i , $i \in \mathcal{J}^- \cup \mathcal{J}^0$ such that*

$$L'_x(x_*, u) = 0, \quad u^i \geq 0, \quad i \in \mathcal{J}^-, \quad u^i f_i(x_*) = 0, \quad i \in \mathcal{J}^-$$

and

$$(L''_{xx}(x_*, u) p, p) \geq 0$$

for all p which satisfy inequalities (4.27).

5. SOME ADDITIONAL INFORMATION

The Newton-Leibnitz formula which establishes the connection between a scalar function $f(x)$ and its derivative is treated in mathematical analysis. This formula is generalized and applies to operator functions.

If $F(x)$ is a differentiable operator function defined in an open convex set $\Omega \in E^n$ and $x, x + h \in \Omega$, then

$$F(x + h) - F(x) = \int_0^1 F'(x + \alpha h) h d\alpha. \quad (5.1)$$

The proof of the formula (which is valid also for operators, defined in functional spaces) can be found, for instance, in the book by A. N. Kolmogorov and S. V. Fomin.

Let us state one more property of operator functions.

If $F(x)$ is a nonlinear differentiable operator function, then for any $x, h, y \in E^n$ the following formula is valid:

$$(F(x + h) - F(x), y) = (F'(x + \theta h) h, y), \quad 0 \leq \theta \leq 1. \quad (5.2)$$

This formula is called Lagrange's formula for operators (or Lagrange's generalized formula). Its proof (for operators of a more general form) can be found in M. M. Vainberg's monograph [1].

In the following chapters we shall have many occasions of using Taylor's formula with the remainder term in Lagrange's form.

If $f(x)$ is a twice continuously differentiable function in a convex set Ω , then for any $x, x + h \in \Omega$ and $\alpha \in [0, 1]$

$$f(x + \alpha h) - f(x) = \alpha (f'(x + \alpha \theta_1 h), h)$$

and

$$f(x + \alpha h) = f(x) + \alpha (f'(x), h) + \frac{\alpha^2}{2} (f''(x + \alpha \theta_2 h) h, h)$$

where $\theta_1, \theta_2 \in [0, 1]$.

Bibliographic Notes

The properties of convex sets and convex functions in finite-dimensional spaces are described by S. Karlin, G. Zoutendijk [1], H. Künzi and W. Krelle. The most comprehensive description is given in R. T. Rockafellar's book. The properties of convex sets and functions in functional spaces are in detail investigated by N. Dunford and J. T. Schwartz.

Special properties connected with strict and strong convexities are considered by E. S. Levitin and B. T. Polyak.

Among many works devoted to the theory of necessary conditions for extrema we note the books by S. Karlin, G. Zoutendijk [1], S. I. Zukhovitskii and L. I. Avdeyeva, H. Künzi and W. Krelle which consider the problems of linear and convex programming in finite-dimensional spaces. A more complete theory

of the necessary conditions of the first order in the general case the reader can find in the works by A. Ya. Dubovitsky and A. A. Milyutin, L. W. Neustadt, H. Halkin and L. W. Neustadt, B. N. Pshenichny [1], M. R. Hestenes [2], K. Arrow, L. B. Hurwitz and H. Uzawa. M. R. Hestenes considers also the necessary conditions of the second order.

The problems of linear programming as well as computational algorithms and the theory of duality in linear programming are treated in detail by D. Gale, G. B. Dantzig [1], S. Karlin, S. I. Zukhovitskii and L. I. Avdeyeva. The general case of the theory of duality in convex programming is treated by E. G. Golstein [1], [2] and R. T. Rockafellar.

The theorem on implicit functions, which is used in deducing the necessary conditions for extrema, can be found in G. M. Fikhtenhol'ts' book.

CHAPTER II

METHODS OF UNCONSTRAINED FUNCTION MINIMIZATION

This chapter is devoted to the problem of minimization of the function $f(x)$ defined in an n -dimensional Euclidean space E^n . Accordingly, in this chapter x is always an n -dimensional vector.

In solving the problem we shall use iterative processes of the type

$$x_{k+1} = x_k + \alpha_k p_k \quad (0.1)$$

where p_k is a vector determining the direction of motion from point x_k and α_k is a numerical factor whose value determines the length of the step in the direction of p_k .

The process (0.1) will be defined if the methods of constructing vector p_k and computing the value of α_k are given for every iteration. The properties of the process—the values of the function for different elements of the sequence $\{x_k\}$, convergence of the sequence to the solution, the rate of convergence, etc.—depend directly on the method chosen. At the same time various methods of constructing vector p_k and determining α_k require different amounts of calculations and involve different constraints on the function to be minimized.

Let us state the considerations on which we shall base our choice of the direction of motion and the step length.

In order to get nearer to point x_* (in the general case x_* is the point at which the necessary conditions for an extremum of function $f(x)$ are satisfied, possibly within a certain accuracy), one should naturally move from point x_k in the direction in which the function decreases, i.e. in the direction of descent. If point x_k is not the point of minimum or a stationary point, then there is an infinite number of vectors p which determine the direction of descent from point x_k and each vector is defined by

$$(f'(x_k), p) < 0$$

($f(x)$ is differentiable).

This is seen from the following argument.

Let $x = x_k + \alpha p$. Expansion of the function in Taylor's series about x_k (it is obviously assumed that the function is differentiable and has a sufficient number of derivatives) gives

$$f(x) = f(x_k) + \alpha (f'_k, p) + \frac{\alpha^2}{2} (f''_{kc} p, p).$$

We set here $f'_k = f'(x_k)$, $f''_{kc} = f''(x_{kc})$, $x_{kc} = x_k + \theta(x - x_k)$, $\theta \in [0, 1]$. For the sake of brevity, these notations will often be used further on in this chapter.

If $(f'_k, p) < 0$ then at least with small values of α , $f(x) < f(x_k)$ since the sign of the right-hand side is determined by a term which is linear with respect to α .

By applying various methods in choosing the direction of the descent and factor α_k , we can obtain different minimization algorithms.

1. GRADIENT METHODS

Method of Steepest Descent

The simplest approach to the choice of the direction of p_k in order to satisfy the condition $(f'_k, p_k) < 0$ (i.e. of the direction of the descent of $f(x)$) is to assume $p_k = -f'_k$.

The iterative process

$$x_{k+1} = x_k - \alpha_k f'(x_k), \quad \alpha_k > 0, \quad k = 0, 1, \dots \quad (1.1)$$

which results from such a choice of the direction of motion is called the *method of steepest descent* or *gradient method*.

In terms of coordinates, process (1.1) is written down in the following form:

$$x_{k+1}^i = x_k^i - \alpha_k \frac{\partial f(x_k)}{\partial x^i}, \quad i = 1, 2, \dots, n.$$

At present the method of steepest descent is one of the best known minimization methods.

The popularity of the method has been favoured by its being comparatively simple and suitable for application to the minimization of a very broad class of functions.

We turn now to the study of the properties of the algorithm (1.1). First of all we describe the method of choosing the magnitude of the scalar factor α_k .

(1) Take an arbitrary value of α (the same at all iterations) and determine point $x = x_k - \alpha f'_k$.

(2) Compute $f(x) = f(x_k - \alpha f'_k)$.

(3) Verify the inequality

$$f(x) - f(x_k) \leq \varepsilon \alpha (f'_k, p_k) \quad (1.2)$$

where $0 < \varepsilon < 1$ is an arbitrarily chosen constant (the same with any $k = 0, 1, \dots$).

(4) If inequality (1.2) is satisfied, then the value of α is taken to be the sought one: $\alpha_k = \alpha$. However if the inequality is not satisfied, we reduce α (multiplying α by an arbitrary $\lambda < 1$) until inequality (1.2) is satisfied.

The above method of choosing α_k needs substantiation; the conditions of the existence of nonzero values of α which satisfy inequality (1.2) must be established. Such a substantiation is given in the following theorem.

Theorem 1.1. *If function $f(x)$ is bounded from below, its gradient $f'(x)$ satisfies Lipschitz' condition*

$$\|f'(x) - f'(y)\| \leq R \|x - y\| \quad (1.3)$$

with any $x, y \in E^n$ and the choice of the value of α_k is made as described above, then in the process (1.1) $\|f'_k\| \rightarrow 0$ as $k \rightarrow \infty$ whatever the initial point x_0 .

Proof. By the mean-value theorem we have

$$f(x) - f(x_k) = (f'(x_{kc}), x - x_k) \quad (1.4)$$

where $x_{kc} = x_k + \theta(x - x_k)$, $\theta \in [0, 1]$. We note that in what follows the index kc (c) will be used to denote an intermediate point in a corresponding segment.

Equality (1.4) can be transformed as follows

$$f(x) - f_k = (f'_k, x - x_k) + (f'_{kc} - f'_k, x - x_k).$$

Hence, noticing that $x - x_k = -\alpha f'_k$ and using (1.3) we obtain

$$\begin{aligned} f(x) - f_k &\leq -\alpha (f'_k, f'_k) + \alpha R \|x_{kc} - x_k\| \|f'_k\| \\ &\leq -\alpha \|f'_k\|^2 + \alpha R \|x - x_k\| \|f'_k\| = \alpha \|f'_k\|^2 (-1 + \alpha R). \end{aligned}$$

The estimate obtained shows that there are values $\alpha \neq 0$ such that the inequality (1.2) is satisfied; to obtain this result it suffices to choose α such that $-1 + \alpha R \leq -\varepsilon$. This is always feasible, since R is a limited quantity and $0 < \varepsilon < 1$. Consequently, (1.2) will always be satisfied with $\alpha \leq \frac{1-\varepsilon}{R}$. Thus, choosing α_k in accordance with the above algorithm we obtain

$$f_{k+1} - f_k \leq -\varepsilon \alpha_k \|f'_k\|^2, \quad (1.5)$$

i.e. with any k we have $f_{k+1} - f_k < 0$ (provided $\|f'_k\| \neq 0$). Since by hypothesis the function has a lower bound, the last inequality

gives as $k \rightarrow \infty$

$$f_{k+1} - f_k \rightarrow 0. \quad (1.6)$$

It follows from (1.5) that

$$\|f'_k\|^2 \leq \frac{f_k - f_{k+1}}{\varepsilon \alpha_k}. \quad (1.7)$$

We note that this algorithm for choosing α_k ensures that $\alpha_k \geq \bar{\alpha} > 0$, with any k , where $\bar{\alpha}$ can be any constant which does not exceed the quantity $\frac{1-\varepsilon}{R}$, since as it was mentioned, the inequality (1.2) or (1.5) is certainly satisfied with $\alpha = \frac{1-\varepsilon}{R}$. With this remark, conditions (1.6) and (1.7) imply that $\|f'_k\| \rightarrow 0$ as $k \rightarrow \infty$, and this proves the theorem.

The class of functions satisfying the requirements of theorem 1.1 is very broad. Such functions can have no minimum point at all, can have local minima, saddle points, etc. Theorem 1.1 shows that the gradient method provides for convergence either to the exact lower bound $\inf_x f(x)$ or to a value of the function at a certain stationary point. The convergence of the sequence $\{x_k\}$ to a stationary point (if such a point exists) also takes place. It is difficult however to determine the rate of convergence with the conditions of theorem 1.1. If the requirements concerning the smoothness and convexity of the function are sufficiently strict, then not only can the convergence of the sequence $\{x_k\}$ be proved but the rate of convergence can also be estimated.

Theorem 1.2. *Let $f(x)$ be a twice continuously differentiable function, and the matrix of second derivatives satisfies the conditions*

$$m \|y\|^2 \leq (f''(x)y, y) \leq M \|y\|^2, \quad M \geq m > 0 \quad (1.8)$$

with any $x, y \in E^n$, and the sequence $\{x_k\}$ be constructed by method (1.1), where α_k is chosen in the way described above. Then with any initial point x_0 we have $x_k \rightarrow x_*$, $f(x_k) \rightarrow f(x_*)$, where x_* is the (unique) point of minimum of $f(x)$.

Then the following estimates of the rate of convergence hold:

$$f_k - f_* \leq q^k [f_0 - f_*], \quad \|x_k - x_*\| \leq Cq^{k/2}, \\ C < \infty, \quad 0 < q < 1. \quad (1.9)$$

Proof. The existence of the unique minimizer of $f(x)$ with the conditions of the theorem follows from the results of lemma 3.2 (Ch. I). Therefore we have only to prove the convergence of sequence $\{x_k\}$ to point x_* and to obtain estimates (1.9). Let us first establish that the first of estimates (1.9) holds. Using Taylor's formula we

obtain

$$f(x_*) = f(x) + (f'(x), x_* - x) + \frac{1}{2} (f''(x_c)(x_* - x), x_* - x).$$

Hence, applying (1.8) we have

$$\begin{aligned} f(x) - f(x_*) &\leq (f'(x), x - x_*) - \frac{m}{2} \|x - x_*\|^2 \\ &\leq \|f'(x)\| \|x - x_*\| - \frac{m}{2} \|x - x_*\|^2. \end{aligned} \quad (1.10)$$

At the same time (since $f'(x_*) = 0$)

$$f(x) - f(x_*) = \frac{1}{2} (f''(x_{c1})(x - x_*), x - x_*)$$

and therefore, from (1.8), we obtain

$$\frac{m}{2} \|x - x_*\|^2 \leq f(x) - f(x_*) \leq \frac{M}{2} \|x - x_*\|^2. \quad (1.11)$$

We have from the left-hand side of inequality (1.11) and from (1.10)

$$\|x - x_*\| \leq \frac{\|f'(x)\|}{m}, \quad (1.12)$$

and from the right-hand side of inequality (1.11)

$$\|x - x_*\|^2 \geq \frac{2}{M} [f(x) - f(x_*)].$$

With the estimates obtained we can write (1.10) in the form

$$f(x) - f(x_*) \leq \frac{\|f'(x)\|^2}{m} - \frac{m}{M} [f(x) - f(x_*)].$$

Hence

$$\|f'(x)\|^2 \geq m \left(1 + \frac{m}{M}\right) [f(x) - f(x_*)]. \quad (1.13)$$

Applying this estimate we can obtain from inequality (1.5)

$$f_{k+1} - f_k \leq -\varepsilon \alpha_k m \left(1 + \frac{m}{M}\right) (f_k - f_*). \quad (1.14)$$

With the conditions of the theorem we have

$$\begin{aligned} f(x) - f(x_k) &= (f'_k, x - x_k) + \frac{1}{2} (f''_{kc}(x - x_k), x - x_k) \\ &= -\alpha \|f'_k\|^2 + \frac{\alpha^2}{2} (f''_{kc} f'_k, f'_k) \leq -\alpha \left(1 - \frac{\alpha M}{2}\right) \|f'_k\|^2. \end{aligned}$$

It follows that the inequality (1.2) is always satisfied if $1 - \frac{\alpha M}{2} \geq \varepsilon$, i.e. if $\alpha \leq \bar{\alpha} = \frac{2(1-\varepsilon)}{M}$. Then we have from (1.14)

$$f_{k+1} - f_* \leq \left[1 - \varepsilon \alpha_k m \left(1 + \frac{m}{M}\right)\right] (f_k - f_*) \leq q (f_k - f_*)$$

where $q = 1 - \bar{\varepsilon} \bar{\alpha} m \left(1 + \frac{m}{M}\right) < 1$, i. e.

$$(f_k - f_*) \leq q^k (f_0 - f_*). \quad (1.15)$$

Since $\bar{\alpha} = \frac{2(1-\varepsilon)}{M}$ we have

$$q = 1 - \frac{2\varepsilon(1-\varepsilon)m}{M} \left(1 + \frac{m}{M}\right).$$

Hence the minimum value of the ratio of the progression $q_{\min}$ is attained with $\varepsilon = \frac{1}{2}$ and then

$$q_{\min} = 1 - \frac{m}{2M} \left(1 + \frac{m}{M}\right).$$

Consequently it is expedient to take $\varepsilon = \frac{1}{2}$ in the condition (1.2).

Estimate (1.15) together with the left-hand side one of estimates (1.11) makes it possible to establish the convergence and estimate the rate of convergence of the sequence $\{x_k\}$ to the point of minimum:

$$\|x_k - x_*\| \leq \left(\frac{2}{m}\right)^{1/2} (f_k - f_*)^{1/2} \leq \left(\frac{2}{m}\right)^{1/2} (f_0 - f_*)^{1/2} q^{k/2} \leq C q^{k/2}.$$

The theorem is proved.

Analyzing the above proof we see that in order to obtain estimate (1.15) we used in the end only conditions (1.2) and (1.13). We conclude that the class of functions for which estimate (1.15) holds is actually much broader than the class of functions which satisfy conditions (1.8), viz. estimate (1.15) is valid for all functions which satisfy the conditions of theorem 1.1 and moreover the condition

$$\|f'(x)\|^2 \geq \delta [f(x) - f_*], \quad \delta > 0.$$

The proof of the validity of estimate (1.15) in this case is not really connected with the existence of a minimum; one can suppose that $f_* = \inf f(x)$ without trying to establish whether the precise lower bound is attained. It should be stressed however that functions of this class do have a minimum—not necessarily the only one—and that sequence $\{x_k\}$ converges to a certain point x_* , the second of estimate (1.9) holding true for the rate of convergence.

Indeed from (1.1) and (1.7) we have

$$\begin{aligned} \|x_{k+1} - x_k\|^2 &= \alpha_k^2 \|f'_k\|^2 \leq \frac{\alpha_k}{\varepsilon} (f_k - f_{k+1}) \\ &\leq \frac{\alpha_k}{\varepsilon} (f_k - f_*) \leq \frac{\alpha_{\max}}{\varepsilon} q^k [f_0 - f_*] \leq C_1 q^k \end{aligned}$$

where $\alpha_{\max}$ is the maximum value of the parameter at which we

start to reduce α . Taking this into account we obtain for any $m > k$

$$\|x_m - x_k\| \leq \sum_{i=k}^{m-1} \|x_{i+1} - x_i\| C_1^{1/2} \sum_{i=k}^{m-1} q^{i/2} \leq C_1^{1/2} \frac{q^{h/2}}{1 - q^{1/2}}.$$

Hence $\|x_m - x_k\| \rightarrow 0$ with $k \rightarrow \infty$, i.e. sequence $\{x_k\}$ converges (to a certain minimum x_*), and also

$$\|x_* - x_k\| = \lim_{m \rightarrow \infty} \|x_m - x_k\| \leq C_1^{1/2} \frac{q^{h/2}}{1 - q^{1/2}} = C_2 q^{h/2}.$$

Variants of the Method

The method of choosing α_k in process (1.1) which involves the checking of inequality (1.2) as described above is not the only one possible. We shall now consider several other methods of choosing the value of α_k ; each of these methods determines a different variant of the gradient method. In proving theorems 1.1 and 1.2 it was established that inequality (1.2) is always satisfied with values of $\alpha \leq \frac{1-\varepsilon}{R}$ (theorem 1.1) or $\alpha \leq \frac{2(1-\varepsilon)}{M}$ (theorem 1.2). This circumstance made it possible to prove the statements about the properties of method (1.1) in choosing α_k under the condition of satisfying inequality (1.2). If constants R or M are known which characterize the function $f(x)$ being minimized, then in applying method (1.1) we can beforehand choose $\alpha_k \equiv \bar{\alpha}$, where $0 < \bar{\alpha} \leq \frac{1-\varepsilon}{R}$ or $0 < \bar{\alpha} \leq \frac{2(1-\varepsilon)}{M}$, and theorems 1.1 and 1.2 will remain valid. This variant of the gradient method allows to determine more exactly the value of the ratio q in the estimates of the rate of convergence (1.9).

Theorem 1.3. *If function $f(x)$ satisfies the conditions of theorem 1.2 and in method (1.1) $\alpha_k = \alpha$, $0 < \alpha < \frac{2}{M}$, then for the rate of convergence of sequence $\{x_k\}$ the estimate*

$$\|x_k - x_*\| \leq q^k \|x_0 - x_*\|,$$

$$q = \max \{ |1 - \alpha m|, |1 - \alpha M| \}$$

is valid, the minimum value, $q_{\min} = \frac{M-m}{M+m}$, is attained with $\alpha = \frac{2}{M+m}$.

Proof. We have

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &= (x_k - \alpha f'_k - x_*, x_{k+1} - x_*) \\ &= (x_k - x_* - \alpha(f'_k - f'_*), x_{k+1} - x_*). \end{aligned}$$

Applying Lagrange's formula for operators (5.2) (Chap. I) we obtain

$$(\alpha(f'_k - f'_*), x_{k+1} - x_*) = (\alpha f''_{kc}(x_k - x_*), x_{k+1} - x_*).$$

Using this result we have

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &= ((I - \alpha f''_{kc})(x_k - x_*), x_{k+1} - x_*) \\ &\leq \|I - \alpha f''_{kc}\| \|x_k - x_*\| \|x_{k+1} - x_*\|, \end{aligned}$$

i.e.

$$\|x_{k+1} - x_*\| \leq \|I - \alpha f''_{kc}\| \|x_k - x_*\| = q \|x_k - x_*\|.$$

By conditions (1.8)

$$q = \|I - \alpha f''_{kc}\| = \max \{|1 - \alpha m|, |1 - \alpha M|\}.$$

In the interval $[0, \frac{2}{M}]$ the linear function $1 - \alpha M$, as we know, changes its sign. Therefore the minimum value $q_{\min}(\alpha)$ will be attained with $1 - \alpha m = -(1 - \alpha M)$, i.e. with $\alpha = \frac{2}{M+m}$, and obviously $q_{\min} = \frac{M-m}{M+m}$.

The theorem is proved.

Note that with $\alpha = \frac{2}{M+m}$ the first of estimates (1.9) can be defined more exactly

$$f_{k+1} - f_* \leq \left(\frac{M-m}{M+m}\right)^2 (f_k - f_*). \quad (1.16)$$

We describe another method of choosing the step length. One can select the value of α_k providing the minimum of the function in the direction of descent, i.e. the chosen value of α_k must satisfy the condition

$$f(x_k - \alpha_k f'_k) = \min_{\alpha \geq 0} f(x_k - \alpha f'_k). \quad (1.17)$$

With this method of choosing the step length all the above results concerning the properties of method (1.1) remain valid; moreover we obtain more exact bounds on the rate of convergence.

We prove a statement similar to that of theorem 1.1.

Theorem 1.4. *If function $f(x)$ satisfies the requirements of theorem 1.1 and in using method (1.1) α_k is chosen by (1.17), then $\|f'_k\| \rightarrow 0$ as $k \rightarrow \infty$ whatever the initial point x_0 .*

Proof. As in theorem 1.1 we obtain the estimate

$$\begin{aligned} f(x) - f(x_k) &= -\alpha \|f'_k\|^2 - \alpha(f'_{kc} - f'_k, f'_k) \\ &\leq -\alpha \|f'_k\|^2 + \alpha^2 R \|f'_k\|^2. \end{aligned}$$

The minimum of function $\varphi(\alpha) = -\alpha \|f'_k\|^2 + \alpha^2 R \|f'_k\|^2$ is attained with $\alpha_{\min} = \frac{1}{2R}$ and $\varphi(\alpha_{\min}) = -\frac{\|f'_k\|^2}{4R}$. Since $\alpha^2 R \|f'_k\|^2$ is the upper bound of the term $-\alpha(f'_{kc} - f'_k, f'_k)$, the value of α_k which satisfies condition (1.17) is obviously not less than $\alpha_{\min}$ and

$$f_{k+1} - f_k \leq \frac{\|f'_k\|^2}{4R}. \quad (1.18)$$

Hence by the same argument as in theorem 1.1 we find that $\|f'_k\| \rightarrow 0$. Q.E.D.

The estimates (1.9) for the variant of the gradient method where the step length is chosen according to (1.17) can be verified in a way analogous to that used in theorem 1.2 with the only difference that expression (1.13) should be used in the estimate $f_{k+1} - f_k \leq -\frac{1}{2M} \|f'_k\|^2$ obtained by the same argument as for (1.18). However we proceed using the results of theorem 1.3. This ensures higher accuracy for the value of the ratio q .

Set $\tilde{x}_{k+1} = x_k - \frac{2}{M+m} f'_k$; then estimate (1.16)

$$f(\tilde{x}_{k+1}) - f(x_*) \leq \left(\frac{M-m}{M+m}\right)^2 (f_k - f_*)$$

holds.

If point x_{k+1} is chosen by applying the condition for function minimization in the direction of descent, then

$$\begin{aligned} f(x_{k+1}) - f(x_*) &\leq f(\tilde{x}_{k+1}) - f(x_*) \leq \left(\frac{M-m}{M+m}\right)^2 (f_k - f_*) \\ &\leq \left(\frac{M-m}{M+m}\right)^{2(k+1)} (f_0 - f_*). \end{aligned}$$

Using now estimates (1.11) we obtain

$$\|x_{k+1} - x_*\|^2 \leq \frac{2}{m} (f_{k+1} - f_*) \leq \frac{2}{m} \left(\frac{M-m}{M+m}\right)^{2(k+1)} \frac{M}{2} \|x_0 - x_*\|^2$$

and finally

$$\|x_{k+1} - x_*\| \leq C \left(\frac{M-m}{M+m}\right)^{k+1}$$

where

$$C = \left(\frac{M}{m}\right)^{1/2} \|x_0 - x_*\|.$$

Thus the following theorem is valid.

Theorem 1.5. *If function $f(x)$ satisfies the conditions of theorem 1.2 and in applying method (1.1) α_k is chosen according to condition (1.17),*

then sequence $\{x_k\}$ converges to the minimum at the rate of a geometric progression whose ratio $q = \frac{M-m}{M+m}$.

Note that the variant of method (1.1) where the step length is chosen according to the condition of function minimization in the direction of descent is often called in literature the method of steepest descent.

Other Gradient Methods

Let $F(x)$ be an arbitrary symmetrical matrix which satisfies the conditions

$$\rho \|y\|^2 \leq (F(x)y, y) \leq R \|y\|^2, \quad \rho > 0 \quad (1.19)$$

with any $x, y \in E^n$. If we choose vector $p = -F(x)f'(x)$, then $(f'(x), p) = -(f', Ff') \leq -\rho \|f'\|^2 < 0$ provided $\|f'(x)\| \neq 0$. Thus vector $p = -F(x)f'(x)$ determines the direction of the decrease of function $f(x)$. Then in order to minimize $f(x)$ we can construct the iterative process

$$x_{k+1} = x_k - \alpha_k F_k f'(x_k), \quad \alpha_k > 0, \quad k = 0, 1, \dots$$

where $\{F_k\}$ is a sequence of arbitrary matrices which satisfy conditions (1.19). To connect this process with the descriptions of the following part of this chapter (to be more precise to make consistent the notations used), we shall consider the process

$$x_{k+1} = x_k - \alpha_k F_k^{-1} f'_k, \quad \alpha_k > 0 \quad (1.20)$$

in which the matrix is the reciprocal of matrix F_k . This affects the heart of the matter in nothing for if matrix F_k satisfies conditions (1.19), then for matrix F_k^{-1} the conditions (cf. lemma 2.9 of Ch. I)

$$m_1 \|y\|^2 \leq (F^{-1}y, y) \leq M_1 \|y\|^2, \quad m_1 = \frac{\rho}{R^2} > 0, \quad M_1 = \frac{1}{\rho} \quad (1.21)$$

will be satisfied and therefore

$$(f'_k, p_k) = -(f'_k, F_k^{-1} f'_k) \leq -m_1 \|f'_k\|^2 \leq 0. \quad (1.22)$$

Different iterative processes will correspond to different sequences $\{F_k^{-1}\}$.

As far as the principles of the methods are concerned the study of method (1.20) does not involve any new elements as compared to the "pure" gradient method (1.1). All the results obtained for method (1.1) remain valid also for method (1.20) with the same requirements to the function being minimized and the same methods of choosing the step length. Only the technique of proving the corresponding statements is lightly changed. Of course, the quantitative values of the parameters in (1.20) will differ from the values of analogous

parameters in method (1.1). In particular, this relates to the value of the ratio q in the estimates of the rate of convergence.

We shall dwell now only on the results of method (1.20) which will be made use of later on.

Theorem 1.6. *The results of theorem 1.2 remain valid for method (1.20).*

Proof. If $x = x_k + \alpha p_k$, where $p_k = -F_k^{-1}f'_k$ then

$$\begin{aligned} f(x) - f(x_k) &= \alpha (f'_k, p_k) + \frac{\alpha^2}{2} (f''_k p_k, p_k) \\ &\leq \alpha (f'_k, p_k) \left(1 + \frac{\alpha M \|p_k\|^2}{2 (f'_k, p_k)} \right). \end{aligned}$$

Now using (1.19) we have

$$(f'_k, p_k) = -(F_k p_k, p_k) \leq -\rho \|p_k\|^2. \quad (1.23)$$

Consequently

$$f(x) - f(x_k) \leq \alpha (f'_k, p_k) \left(1 - \frac{\alpha}{2} \frac{M}{\rho} \right).$$

Hence inequality (1.2) will be satisfied without fail if $1 - \frac{\alpha M}{2\rho} \geq \varepsilon$, i.e. if $\alpha \leq \bar{\alpha} = \frac{2(1-\varepsilon)}{M} \rho$. This substantiates the method of choosing α_k .

Since $(f'_k, p_k) < 0$ with $\|f'_k\| \neq 0$, it follows from the condition

$$f_{k+1} - f_k \leq \varepsilon \alpha_k (f'_k, p_k) \quad (1.24)$$

that $f_{k+1} < f_k$. Using now (1.24) and having in mind that $f(x)$ is bounded from below, by analogy to the proof in theorem 1.1 that $\|f'_k\| \rightarrow 0$, we establish that $(f'_k, p_k) \rightarrow 0$ as $k \rightarrow \infty$. By (1.22), it means that $\|f'_k\| \rightarrow 0$. Hence, since $f(x)$ is strongly convex, sequence (1.20) converges to the solution x_* . In order to obtain bounds on the rate of convergence of $f_k \rightarrow f_*$, $x_k \rightarrow x_*$, let us write inequality (1.24), using (1.22), in the form $f_{k+1} - f_k \leq -\varepsilon \alpha_k m_1 \|f'_k\|^2$. Further, introducing in this expression $\|f'_k\|$ with the aid of inequality (1.13) and applying the argument of theorem 1.2 we establish the validity for method (1.20) of the estimates of the convergence rate (1.9). The value of the ratio of the progression is

$$q = 1 - \varepsilon \bar{\alpha} m_1 m \left(1 + \frac{m}{M} \right) = 1 - \varepsilon \frac{2(1-\varepsilon)\rho}{M} m_1 m \left(1 + \frac{m}{M} \right).$$

The minimum of q is attained with $\varepsilon = \frac{1}{2}$:

$$q_{\min} = 1 - \frac{\rho^2 m}{2R^2 M} \left(1 + \frac{m}{M} \right).$$

The theorem is proved.

It follows from the proof that process (1.20) remains convergent if we set $\alpha_k = \alpha$, $0 < \alpha < \frac{2}{M} \rho$ (a variant of the method with constant step). By the same argument as in theorem 1.3, one can obtain the estimate

$$\|x_{k+1} - x_*\| \leq \|I - \alpha F_k^{-1} f''_{kc}\| \|x_k - x_*\|.$$

However it is impossible to obtain an estimate of the ratio of the progression as was done in theorem 1.3, since the matrix $F_k^{-1} f''_{kc}$ is not positive definite in the general case (this last property is fulfilled only on condition that matrices F_k^{-1} and $f''(x_{kc})$ can be transposed).

We can consider a variant of method (1.20) in which the step length is chosen using the condition that $f(x)$ attains minimum in the direction of descent.

Theorem 1.7. *If function $f(x)$ satisfies the conditions of theorem 1.2 and in applying method (1.20) parameter α_k is chosen using condition*

$$f(x_k + \alpha_k p_k) = \min_{\alpha \geq 0} f(x_k + \alpha p_k),$$

then sequence $\{x_k\}$ converges to the point of minimum at the rate of a geometric progression.

The theorem can be proved according to the following scheme. Expanding the function into Taylor's series to second-order terms about point x_k and reasoning as in theorem 1.4 we can obtain the estimate:

$$f_{k+1} - f_k \leq -\frac{1}{2} \frac{(f'_k, p_k)^2}{M \|p_k\|^2}.$$

This inequality, because of (1.22) and (1.23), is equivalent to

$$f_{k+1} - f_k \leq -\frac{1}{2} \frac{\rho m_1 \|f'_k\|^2}{M}.$$

Further, expressing $\|f'_k\|$ with the use of inequality (1.13) one should repeat completely the argument of theorem 1.2. We cannot obtain in this case a more precise value of the ratio q since we know it must be greater than in the method of steepest descent.

Qualitative Analysis of the Methods

Let us compare the gradient methods considered above and consider certain statements on the quality of these algorithms, i.e. on their effectiveness in solving minimization problems.

We have studied three variants of method (1.1) differing in the method of choosing the step length. The properties of the variants resemble closely. They can be used in minimizing functions of like

classes, their rate of convergence being nearly the same (in the cases where it can be estimated). Consequently it is expedient in solving problems to use that of the variants of the method which requires the minimum of labour. Computational effort at each iteration in the variants of process (1.1) can evidently be different only owing to the difference in the method of choosing parameter α_k . The variant of the method which uses a constant step $\alpha_k = \alpha$ requires the least amount of computations at iterations (for in this case it is necessary to calculate only the gradient $f'(x_k)$). However in most problems such a method of choosing α_k is practically impossible since usually the values of constants R, M characterizing the function are unknown.

Let us compare the amounts of work required by the methods of choosing the step length; this is connected with the checking of conditions (1.2) and (1.17). We have established that if the function $f(x)$ satisfies certain requirements (theorems 1.1, 1.2), then inequality (1.2) is always satisfied, at least with sufficiently small values of α_k (which are determined by the values of constants R, M). By virtue of this whatever the value $\alpha_{\max}$ from which we began to verify inequality (1.2), the inequality will be satisfied, after a certain finite number of reductions of the parameter, i.e. we have to calculate the function value a finite number of times to choose the required value of α_k . As to the choice of α_k using condition (1.17) it is in the general case a procedure with an infinite number of possible values.

Of course, in practice we have to determine a point of minimum in the direction of descent by evaluating the function also a finite number of times. Clearly for a more or less precise solution of a one-dimensional minimization problem we have to perform more calculations of the function value than for satisfying inequality (1.2). The above considerations show that one should prefer the method of choosing the step length which involves checking of inequality (1.2).

All of the above remarks can be applied to method (1.20) too.

The reader understands surely that the above arguments are based only on the use of the most general properties of the function being minimized and the algorithms being studied and do not make use of the specific properties of concrete functions. Therefore the above recommendations should not be considered as absolutely applicable. This remark should be kept in mind in what follows.

We turn now to the discussion of the effectiveness of gradient methods. From the point of view of solving problems of function minimization, for sufficiently good functions (smooth, convex) gradient methods give convergence to the minimum at the rate of a geometric progression. The value of the ratio of the progression, in particular for strongly convex functions, depends on the maximum M and minimum m of the eigenvalue of the matrix of second derivatives of function $f(x)$. The ratio q will be sufficiently small only when the eigenvalues m and M are but slightly different, i.e. when the matrix

$f''(x)$ is well conditioned. In this case the rate of convergence is fast. However in computational practice such problems occur very seldom. As a rule we have to find the minimum of functions whose matrix $f''(x)$ is ill conditioned ($\frac{m}{M} \ll 1$). The less the ratio $\frac{m}{M}$, the closer to unity will be the ratio q of the progression and the slower the rate of convergence. This fact can be given a geometric interpretation. With the diminishing of the ratio $\frac{m}{M}$, the graph of the function being minimized (i.e. surfaces $f(x) = C$) becomes more elongated and the direction of vector $f'(x)$ at most points deviates more and more from the direction to the point of minimum. This leads to slowing down the rate of convergence. This can be particularly well visualized by considering, for instance, a strictly convex quadratic function $f(x)$ in space E^2 , e.g. $f = \frac{1}{2} \left(\frac{x^2}{a^2} + \frac{y^2}{b^2} \right)$. The matrix of second derivatives of this function has constant elements, its level surfaces are ellipses $\frac{1}{2} \left(\frac{x^2}{a^2} + \frac{y^2}{b^2} \right) = C$, the point of minimum coincides with the centre of the ellipses. The eigenvalues of the matrix of second derivatives are $\frac{1}{a^2}$ and $\frac{1}{b^2}$. The more the ratio $\frac{a^2}{b^2}$ differs from unity, the more are the lines of level extended along one of the axes OX or OY and the greater is the number of steps in the direction of the anti-gradient which have to be taken in moving from an arbitrary point (x_0, y_0) in order to attain a sufficiently small neighbourhood of the minimum point.

The slow convergence of gradient methods prevents their being used in solving complicated minimization problems since too much time is required even with the use of modern high speed computers. Therefore at present minimization methods have been and are being worked out which have a faster rate of convergence, and gradient methods are often used in combination with other more effective ones at the initial stage of solving the problem when point x_k is at a great distance from the minimum and steps along the antigradient permit to obtain a significant decrease of the function. We also stress once more the unmistakable advantages of gradient methods and their suitability for minimizing functions of very different characters.

2. NEWTON'S METHOD WITH STEP ADJUSTMENT

Construction of the Method

In gradient methods only the linear term of the expansion of the function in Taylor's series is used in choosing the direction of motion, i.e. use is made of the crudest approximation to the function being minimized.

Let function $f(x)$ whose minimum is to be determined be strictly convex and sufficiently smooth.

Consider the function

$$\psi(x) = f(y) + (f'(y), x - y) + \frac{1}{2} (f''(y)(x - y), x - y)$$

which is a quadratic approximation to $f(x)$ in the neighbourhood of a certain point y . Since function $f(x)$ is strictly convex, function $\psi(x)$ as can easily be ascertained is also strictly convex; therefore the minimum of this function is attained at a unique point and vector $p = \bar{y} - y$ which minimizes $\psi(x)$ is determined from the formula $p = -(f''(y))^{-1} f'(y)$. The direction determined by vector p is that of descent of $f(x)$ since $(f'(y), p) = -(f''(y)p, p) < 0$ by virtue of $f(x)$ being convex. The quadratic function $\psi(x)$ in the neighbourhood of point y , is a far better approximation to the function being minimized than a linear function. Therefore one naturally expects, at least if point y is in a sufficiently small neighbourhood of solution x_* , that by moving from point y in the direction $p = -(f''(y))^{-1} f'(y)$ one can attain a more significant decrease of the function and obtain a more accurate approximation to the solution than by moving in the direction $-f'(y)$ which is used in the gradient method. On the ground of the above argument we suppose that the iterative process

$$x_{k+1} = x_k - \alpha_k (f''_k)^{-1} f'_k, \quad \alpha_k > 0, \quad k = 0, 1, \dots \quad (2.1)$$

when used to construct successive approximations to the solution of the problem of minimization of function $f(x)$ will prove more effective than the method of steepest descent, i.e. that the rate of convergence of $x_k \rightarrow x_*$, $f(x_k) \rightarrow f(x_*)$ when using algorithm (2.1) will be faster than when applying the gradient method. The results of this section will show that our expectation is justified.

We shall call method (2.1) *Newton's method with adjustment of steps, or generalized Newton method*.

The usual Newton method corresponds to the case when $\alpha_k = 1$.

Denoting the elements of matrix $(f''_k)^{-1}$ by $\varphi_{ij}(x_k)$, $i, j = 1, 2, \dots, n$, where i is the row index, we can write method (2.1)

in its coordinate form:

$$x_{k+1}^i = x_k^i - \alpha_k \sum_{j=1}^n \varphi_{ij} \frac{\partial f(x_k)}{\partial x^j}, \quad i = 1, \dots, n.$$

Note that method (2.1) can also be presented in the following form:

$$f_k'' p_k = -f_k', \quad x_{k+1} = x_k + \alpha_k p_k$$

or in coordinate form

$$\sum_{j=1}^n \frac{\partial^2 f(x_k)}{\partial x^i \partial x^j} p_k^j = -\frac{\partial f(x_k)}{\partial x^i}$$

$$x_{k+1}^i = x_k^i + \alpha_k p_k^i, \quad i = 1, \dots, n.$$

Consequently, in order to determine vector p_k one can solve a system of linear equations instead of inverting the matrix $f''(x_k)$.

We shall study two variants of the generalized Newton method in which different methods of choosing parameter α will be used. The first of these methods consists of the following four steps:

- (1) Setting $\alpha = 1$ calculate point $x = x_k + \alpha p_k$.
- (2) Evaluate $f(x) = f(x_k + \alpha p_k)$.
- (3) Check the inequality

$$f(x) - f(x_k) \leq \varepsilon \alpha (f_k', p_k), \quad 0 < \varepsilon < \frac{1}{2}. \quad (2.2)$$

- (4) If the inequality is satisfied, then take the value $\alpha = 1$ to be the sought one: $\alpha_k = 1$. Otherwise proceed to reduce α until inequality (2.2) is satisfied.

We shall call further on the above method of choosing the value of α_k *method of choosing α_k according to condition (2.2)*. It can be seen that this method of choosing the step length is analogous to that in the method of steepest descent, involving the checking of inequality (1.2).

The other variant of method (2.1) requires that the value of α_k provide the minimum of the function in the direction of motion

$$f(x_k - \alpha_k (f_k'')^{-1} f_k') = \min_{\alpha \geq 0} f(x_k - \alpha (f_k'')^{-1} f_k'). \quad (2.3)$$

Theorems about Properties of the Method

As follows from formula (2.1) Newton's method can only be applied to the minimization of functions that have an invertible matrix of second derivatives and this matrix $(f_k'')^{-1}$ as will be shown later on must be bounded.

Strongly convex twice continuously differentiable functions possess such properties. Therefore in this section we always assume the function $f(x)$ to satisfy the following conditions:

$$m\|y\|^2 \leq (f''(x)y, y) \leq M\|y\|^2, \quad m > 0 \quad (2.4)$$

for any $x, y \in E^n$. Recall that such functions have a lower bound and a unique minimum point x_* .

Theorem 2.1. *If in minimizing function $f(x)$ that satisfies conditions (2.4) use is made of method (2.1) and α_k is chosen according to condition (2.2), then whatever the initial point x_0 the sequence $\{x_k\}$ converges to the minimum point at a superlinear rate*

$$\|x_{N+l} - x_*\| \leq C\lambda_N \dots \lambda_{N+l} \quad (2.5)$$

where, $N, C < \infty$, $\lambda_{N+l} < 1$ with any $l \geq 0$, $\lambda_i \rightarrow 0$ as $i \rightarrow \infty$.

Proof. Method (2.1) can be considered as a process of the gradient type (1.20), assuming $F_k^{-1} = (f_k'')^{-1}$. Since matrix f_k'' possesses the required properties, the convergence of method (2.1) to the solution follows from the general results pertaining to the convergence of gradient methods (theorem 1.6).

Let us establish the validity of estimate (2.5). Note first of all that

$$(f_k', p_k) = - (f_k'' p_k, p_k) \leq -m\|p_k\|^2. \quad (2.6)$$

Since $(f_k', p_k) < 0$ and $(f_k', p_k) \rightarrow 0$ (theorem 1.6), it follows from (2.6) that $\|p_k\| \rightarrow 0$ as $k \rightarrow \infty$. Let us show now that from a certain iteration on in method (2.1) we have $\alpha_k \equiv 1$. Using Taylor's formula and (2.6) we obtain

$$\begin{aligned} f_{k+1} - f_k &= \alpha_k (f_k', p_k) + \frac{\alpha_k^2}{2} (f_k'' p_k, p_k) + \frac{\alpha_k^3}{2} ((f_{kc}'' - f_k'') p_k, p_k) \\ &\leq \alpha_k (f_k', p_k) \left(1 - \frac{\alpha_k}{2} - \frac{\alpha_k}{2} \frac{\|f_{kc}'' - f_k''\| \|p_k\|^2}{m \|p_k\|^2} \right) \end{aligned}$$

where $x_{kc} = x_k + \theta(x_{k+1} - x_k)$, $\theta \in [0, 1]$. Since $\|x_k - x_*\| \rightarrow 0$, we have as $k \rightarrow \infty$

$$\|f_{kc}'' - f_k''\| \leq \|f_{kc}'' - f_*''\| + \|f_*'' - f_k''\| \rightarrow 0$$

by virtue of the (operator) function $f''(x)$ being continuous. Then for any constant $0 < \varepsilon < \frac{1}{2}$ there is a number $N_0(\varepsilon)$ such that with $k \geq N_0(\varepsilon)$ the condition

$$1 - \frac{\alpha_k}{2} - \frac{\alpha_k}{2} \frac{\|f_{kc}'' - f_k''\|}{m} \geq \varepsilon$$

will be satisfied with $\alpha_k = 1$. This means that inequality (2.2) will also be satisfied with $\alpha_k = 1$. Thus the method used in choosing

the step length under the conditions of the theorem guarantees that method (2.1) from a certain iteration on will be applied with a step equal to unity, i.e. will be transformed into the usual Newton method. We can now obtain bounds on the rate of convergence of the method:

$$(x_{k+1} - x_*, x_{k+1} - x_*) = (x_k - x_* - (f_k'')^{-1} f_k', x_{k+1} - x_*).$$

By Lagrange's formula for operators we have

$$\begin{aligned} ((f_k'')^{-1} f_k', x_{k+1} - x_*) &= ((f_k'')^{-1} (f_k' - f_*'), x_{k+1} - x_*) \\ &= ((f_k'')^{-1} f_{kc}'' (x_k - x_*), x_{k+1} - x_*) \end{aligned}$$

where $x_{kc} = x_k + \theta (x_k - x_*)$, $\theta \in [0, 1]$. Consequently

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &= ((I - (f_k'')^{-1} f_{kc}'') (x_k - x_*), x_{k+1} - x_*) \\ &= ((f_k'')^{-1} (f_k'' - f_{kc}'') (x_k - x_*), x_{k+1} - x_*) \\ &\leq \frac{1}{m} \|f_k'' - f_{kc}''\| \|x_k - x_*\| \|x_{k+1} - x_*\| \end{aligned}$$

or

$$\|x_{k+1} - x_*\| \leq \lambda_k \|x_k - x_*\| \quad (2.7)$$

where $\lambda_k = \frac{1}{m} \|f_k'' - f_{kc}''\|$. Since $\|f_k'' - f_{kc}''\| \rightarrow 0$, there is a number N such that with $k = N + l$, $l = 0, 1, \dots$ we have $\lambda_{N+l} < 1$ and $\lambda_{N+l} \rightarrow 0$ as $l \rightarrow \infty$. Setting $\|x_N - x_*\| = C$ and taking into account the above remarks we obtain estimate (2.5).

The theorem is proved.

Let us suppose now that matrix $f''(x)$ satisfies besides conditions (2.4) also Lipschitz' condition

$$\|f''(x) - f''(y)\| \leq R \|x - y\|, \quad x, y \in E^n. \quad (2.8)$$

In this case we have in estimate (2.7)

$$\lambda_k = \frac{1}{m} \|f_k'' - f_{kc}''\| \leq \frac{R}{m} \|x_k - x_*\|$$

and therefore

$$\|x_{k+1} - x_*\| \leq \frac{R}{m} \|x_k - x_*\|^2. \quad (2.9)$$

Consequently the following theorem is valid.

Theorem 2.2. *If function $f(x)$ is such that conditions (2.4) and (2.8) are satisfied, then sequence (2.1) in which the values of α_k are chosen according to condition (2.2), whatever the initial point x_0 , converges to the solution at a quadratic rate, i.e. estimate (2.9) is valid.*

Estimate (2.9) can be written also in the following form. We set $\mu_k = \frac{R}{m} \|x_k - x_*\|$. There is a number L such that with $k = L + l$, $l = 0, 1, \dots$ we have $\mu_{L+l} < 1$ and

$$\frac{R}{m} \|x_{L+l} - x_*\| \leq \left(\frac{R}{m} \|x_{L+l-1} - x_*\| \right)^2 \leq \dots \leq \left(\frac{R}{m} \|x_L - x_*\| \right)^{2^l}.$$

Finally we can write

$$\|x_{L+l} - x_*\| \leq \frac{m}{R} \mu_L^{2^l}.$$

Let us consider now the variant of method (2.1) with the step length being chosen according to condition (2.3). The convergence of sequence $\{x_k\}$ to the solution in this case follows from the general results about the convergence of gradient methods (theorem 1.7). The rate of convergence as in the case of choosing α_k according to condition (2.2) will be superlinear if condition (2.4) is satisfied and quadratic if condition (2.8) is also satisfied. This can be proved as follows.

Let $\tilde{x}_{k+1} = x_k - (f_k'')^{-1} f_k'$ and $x_{k+1} = x_k - \alpha_k (f_k'')^{-1} f_k'$ where α_k is chosen according to condition (2.3). Then using estimate (1.11) we obtain

$$\frac{m}{2} \|x_{k+1} - x_*\|^2 \leq f_{k+1} - f_* \leq f(\tilde{x}_{k+1}) - f_* \leq \frac{M}{2} \|\tilde{x}_{k+1} - x_*\|^2.$$

By (2.7), $\|\tilde{x}_{k+1} - x_*\| \leq \lambda_k \|x_k - x_*\|$, $\lambda_k \rightarrow 0$ as $k \rightarrow \infty$. Consequently if conditions (2.4) are satisfied, then

$$\|x_{k+1} - x_*\| \leq \left(\frac{M}{m} \right)^{1/2} \lambda_k \|x_k - x_*\| = \gamma_k \|x_k - x_*\| \quad (2.10)$$

where $\gamma_k = \left(\frac{M}{m} \right)^{1/2} \lambda_k \rightarrow 0$ as $k \rightarrow \infty$. If estimate (2.8) holds, then

$$\lambda_k \leq \frac{R}{m} \|x_k - x_*\|, \quad \|x_{k+1} - x_*\| \leq \left(\frac{M}{m} \right)^{1/2} \frac{R}{m} \|x_k - x_*\|^2. \quad (2.11)$$

Modifications of the Generalized Newton Method

As one of the possible modifications of method (2.1) we shall consider an algorithm in which the sequence of approximations to the solution is constructed by the following formula:

$$x_{k+1} = x_k - \alpha_k (f_0'')^{-1} f_k', \quad \alpha_k \geq 0. \quad (2.12)$$

In this method $p_k = -(f_0'')^{-1} f_k'$, i.e. in order to determine the directions of descent use is made of the same matrix $(f_0'')^{-1}$. Method (2.12) is a particular case of algorithm (1.20) ($F_k^{-1} = (f_0'')^{-1}$). Therefore it can be asserted that sequence (2.12), whatever the initial point x_0 ,

will converge to the solution at the rate of a geometric progression both if the step length is chosen according to condition (2.2) and if it is chosen using condition (2.3) (theorems 1.6 and 1.7). However the value of the ratio q , i.e. the actual rate of convergence, will significantly depend on the initial approximation chosen, x_0 .

In fact, taking into account that in method (2.12) $(f'_k, p_k) = -(f'_k, f''_0 f'_k)$ and using Taylor's formula we can obtain the following bounds on the rate of convergence (as in theorem 2.1):

$$f_{k+1} - f_k \leq \alpha_k (f'_k, p_k) \left(1 - \frac{\alpha_k}{2} - \frac{\alpha_k}{2} \frac{\|f''_{kc} - f''_0\|}{m} \right) \quad (2.13)$$

where $x_{kc} = x_k + \theta (x_{k+1} - x_k)$, $\theta \in [0, 1]$. If $x_0 \rightarrow x_*$, then since the matrix of second derivatives is continuous we have

$$\max_{x \in S} \|f''(x) - f''(x_0)\| \rightarrow 0$$

($S = \{x: f(x) \leq f(x_0)\}$). Thus the closer is the initial point x_0 to point x_* , the greater will be the value of α_k which satisfy the inequality (2.2), i.e. the greater will be the step in the process (2.12) if the step length is chosen according to condition (2.2). In particular for any constant $0 < \varepsilon < \frac{1}{2}$ there is a constant $\rho(\varepsilon)$ such that if the initial approximation x_0 was chosen in a sphere S of radius ρ , we shall have

$$\frac{1}{2} - \frac{1}{2} \frac{\max_{x \in S} \|f''(x) - f''_0\|}{m} \geq \varepsilon.$$

This means, by (2.13), that provided the initial approximation was chosen sufficiently close to point x_* , inequality (2.2) will be satisfied with $\alpha_k = 1$, i.e. process (2.12) will converge with a step equal to unity. Then proceeding as in the proof of theorem 2.1 we obtain the following estimate:

$$\begin{aligned} \|x_{k+1} - x_*\| &\leq \|(f''_0)^{-1}\| \|f''_0 - f''_{kc}\| \|x_k - x_*\| \\ &\leq q \|x_k - x_*\| \end{aligned} \quad (2.14)$$

where $q = \frac{1}{m} \max_{x \in S} \|f''_0 - f''(x)\|$. This shows that the value of the ratio q depends on the choice of the initial point x_0 , the value of q becoming the smaller the closer point x_0 lies to the solution x_* .

For the variant of method (2.12) in which the step length is chosen under the condition of $f(x)$ attaining minimum in the direction of

motion, one can find using inequalities (2.14) and reasoning as in obtaining estimates (2.10) and (2.11) that

$$\|x_{k+1} - x_*\| \leq \left(\frac{M}{m}\right)^{1/2} q \|x_k - x_*\| = \tilde{q} \|x_k - x_*\|$$

where $\tilde{q} = \left(\frac{M}{m}\right)^{1/2} \frac{1}{m} \max_{x \in S} \|f''_0 - f''_{kc}\| \rightarrow 0$ if $x_0 \rightarrow x_*$.

Another possible modification of Newton's method is the following.

Let $k = \xi t + i$, $\xi = 0, 1, \dots$, $i = 0, 1, \dots, t - 1$, $t \geq 1$ be an arbitrary integer. We can construct an iterative process

$$x_{\xi t+i+1} = x_{\xi t+i} - \alpha_{\xi t+i} (f''_{\xi t})^{-1} f'_{\xi t+i}, \quad \alpha_k \geq 0$$

or, with the original notations,

$$x_{k+1} = x_k - \alpha_k (f''_{\xi t})^{-1} f'_k, \quad \alpha_k \geq 0. \quad (2.15)$$

Such a method takes an intermediate position between algorithm (2.1) in which for the construction of vector p_k a new matrix $(f''_k)^{-1}$ is used and algorithm (2.12) in which in determining the direction of motion the same matrix $(f''_0)^{-1}$ is always used. In method (2.15) a new matrix is applied after t steps. This algorithm as well as methods (2.1), (2.12) can be considered to be variants of the gradient method (1.20); therefore its convergence with different ways of choosing the step length follows from theorems 1.6 and 1.7.

We consider the rate of convergence of method (2.15), assuming that the step length is chosen according to condition (2.2) and that conditions (2.4) and (2.8) are valid for function $f(x)$.

Using Taylor's formula we obtain

$$f_{k+1} - f_k \leq \alpha_k (f'_k, p_k) \left(1 - \frac{\alpha_k}{2} - \frac{\alpha_k}{2} \frac{\|f''_{kc} - f''_{\xi t}\|}{m}\right).$$

Because of the convergence of the process we have $\|x_{kc} - x_{\xi t}\| = \|x_{(\xi t+i)c} - x_{\xi t}\| \leq \|x_{\xi t} - x_{\xi t+1}\| + \dots + \|x_{\xi t+i} - x_{\xi t+i+1}\| \rightarrow 0$ as $k \rightarrow \infty$, therefore $\|f''_{kc} - f''_{\xi t}\| \rightarrow 0$. Taking this into account and reasoning as we did in theorem 2.1 we can prove that from a certain iteration on, method (2.15) will converge with a step equal to unity: $\alpha_k = 1$. Then, by theorem 2.2, the following estimate is valid

$$\|x_{\xi t+1} - x_*\| \leq \frac{R}{m} \|x_{\xi t} - x_*\|^2 \quad (2.16)$$

with any $\xi \geq L$, where L is a positive number. Further, proceeding as in theorem 2.1 we can obtain the estimate

$$\begin{aligned} \|x_{\xi t+2} - x_*\| &= \|x_{\xi t+1} - (f''_{\xi t})^{-1} f'_{\xi t+1} - x_*\| \\ &\leq \|(f''_{\xi t})^{-1}\| \|f''_{\xi t} - f''_{(\xi t+1)c}\| \|x_{\xi t+1} - x_*\| \end{aligned}$$

where $x_{(\xi t+1)c} = x_{\xi t+1} + \theta(x_* - x_{\xi t+1})$, $\theta \in [0, 1]$. We have by (2.8) $\|f_{\xi t}'' - f_{(\xi t+1)c}''\| \leq \|f_{\xi t}'' - f_*''\| + \|f_*'' - f_{(\xi t+1)c}''\| \leq R(\|x_{\xi t} - x_*\| + \|x_{\xi t+1} - x_*\|)$. Using estimate (2.16) we now obtain

$$\begin{aligned} \|x_{\xi t+2} - x_*\| &\leq \frac{R}{m} (\|x_{\xi t} - x_*\| + \|x_{\xi t+1} - x_*\|) \|x_{\xi t+1} - x_*\| \\ &\leq \frac{R^2}{m^2} \|x_{\xi t} - x_*\|^3 \left(1 + \frac{R}{m} \|x_{\xi t} - x_*\|\right), \end{aligned}$$

i.e.

$$\|x_{\xi t+2} - x_*\| \leq C_2 \|x_{\xi t} - x_*\|^3, \quad C_2 < \infty.$$

Suppose that with a certain $2 < j < t - 1$ the following estimate is satisfied:

$$\|x_{\xi t+j} - x_*\| \leq C_j \|x_{\xi t} - x_*\|^{j+1}, \quad C_j < \infty.$$

Then we have

$$\begin{aligned} \|x_{\xi t+j+1} - x_*\| &= \|x_{\xi t+j} - (f_{\xi t}'')^{-1} f_{\xi t+j}' - x_*\| \\ &\leq \|(f_{\xi t}'')^{-1}\| \|f_{\xi t}'' - f_{(\xi t+j)c}''\| \|x_{\xi t+j} - x_*\| \\ &\leq \frac{R}{m} (\|x_{\xi t} - x_*\| + \|x_{\xi t+j} - x_*\|) \|x_{\xi t+j} - x_*\| \\ &\leq C_{j+1} \|x_{\xi t} - x_*\|^{j+2}, \quad C_{j+1} = \frac{R}{m} C_j (1 + C_j \|x_{\xi t} - x_*\|^j). \end{aligned}$$

Thus the following bound on the rate of convergence is valid for method (2.15):

$$\|x_{(\xi+1)t} - x_*\| \leq C \|x_{\xi t} - x_*\|^{t+1}. \quad (2.17)$$

This estimate means that sequence $\{x_{\xi t}\}$ converges to the solution at a rate of the order of $t + 1$.

Discussion of the Properties of Newton's Method

We have established that Newton's method with adjustment of steps converges to the solution whatever the initial point x_0 at a rate either superlinear or quadratic depending on the requirements satisfied by function $f(x)$.

The convergence of method (2.1) from any initial approximation on is its essential advantage over the usual Newton method in which the convergence is ensured if the initial approximation is sufficiently good (i.e. sufficiently close to the solution of the problem). Besides, in applying Newton's method the check of the conditions which guarantee that the initial approximation ensures the convergence of the process is in practice difficult to perform, since

it requires such data about the function that are usually unknown (for instance, values of the constants m , M).

The comparison of the two methods of choosing the step length according to condition (2.2) or (2.3) is in favour of the former, for it proves less laborious as to the amount of calculations of function values (in particular, from a certain iteration on the former method requires the function to be evaluated only once since $\alpha_k = 1$) and guarantees the rate of convergence not slower than with the latter method.

If we compare Newton's method and gradient methods as applied to solving problems of convex function minimization, it becomes evident that Newton's method ensures a faster rate of convergence of the sequence of approximations to the solution. Thus if we consider the rate of convergence to mean effectiveness of a method, then our supposition stated at the beginning of this section that Newton's method must be far more effective than the gradient methods is justified. However a more precise meaning of the concept of effectiveness of a method is based on estimating the amount of computation involved when applying a concrete algorithm for the solution of a problem to the required accuracy. Consequently the effectiveness of an algorithm can be estimated by the number of iterations, which are necessary for solving the problem, and the amount of computations at each iteration.

The amount of computations per iteration in Newton's method is as a rule considerably greater than in the gradient methods because of the necessary computations and inversions of the matrices of second derivatives. On the other hand, Newton's method usually involves scores and hundreds of times less iterations than gradient methods; by virtue of this fact Newton's method proves to be considerably more effective.

Nevertheless in many problems the labour per iteration in Newton's method can prove excessively great because of it being necessary to calculate matrices of second derivatives, $f''(x)$ (as a rule in solving extremal problems the greatest difficulty is the calculation of the matrix $f''(x)$ and not its inversion). Such problems will be considered later on. In order to solve the problem in such a case, one can make use of one of the modifications of Newton's method which we have studied. In one of the modifications we have to calculate and invert the matrix $f''(x)$ only once, in the other this is made after a finite number of iterations. If the initial approximation is good enough, then the rate of convergence to the solution will be fast. However, using modifications of Newton's method is not a cardinal solution of the problem of reducing the amount of work required to solve the problem (speaking generally, it can become even greater). Therefore we come to the question of the possibility of constructing minimization methods which would be close to Newton's method as to their

rate of convergence and would require considerably less computations at every iteration.

Several such methods have been worked out; they are based on different ideas. As a rule they prove more effective than Newton's method and this is why they are used more and more at present. The next three sections are devoted to the study of such algorithms.

3. METHODS OF DUAL DIRECTIONS

Considerations on the Choice of Schemes of the Methods

In the preceding section we noted that the main difficulty in applying Newton's method is the necessity of evaluating the matrix of second derivatives of the function being minimized. Consequently algorithms which would be more effective than Newton's method should exclude the calculation of second derivatives, providing however the rate of convergence of Newton's method.

The question arises whether it is possible in constructing the sequence of approximations to the solutions to determine directions p_k which would be close to those in Newton's method, by using for this purpose only the first derivative of the function being minimized.

The first and the second derivatives of $f(x)$ can be related by Taylor's formula for operators (the gradient $f'(x)$ is one):

$$f'(y) - f'(x) = f''(x)(y - x) + \omega(x, y - x) \quad (3.1)$$

where $\|\omega(x, y - x)\| = o(\|y - x\|)$.

The equality (3.1) suggests that if we calculate the derivatives $f'(x)$ at arbitrary but close points $x_1, \dots, x_{n+1}$ and determine the square $n \times n$ -matrix A with the aid of a system of (vector) equations

$$f'(x_{i+1}) - f'(x_i) = A(x_{i+1} - x_i), \quad i = 1, \dots, n \quad (3.2)$$

(assuming of course vectors $x_{i+1} - x_i$, $i = 1, \dots, n$ to be linearly independent), then matrix A must be close to the matrix of second derivatives calculated at any point x_i .

In fact, by (3.1) with any i we have

$$f'(x_{i+1}) - f'(x_i) = f''(x_i)(x_{i+1} - x_i) + \omega(x_i, x_{i+1} - x_i)$$

and therefore using (3.2) we obtain

$$\begin{aligned} A(x_{i+1} - x_i) &= f''(x_i)(x_{i+1} - x_i) + \omega(x_i, x_{i+1} - x_i) \\ i &= 1, \dots, n. \end{aligned}$$

This system of equations can be rewritten in the following form:

$$A(x_{i+1} - x_i) = f_j''(x_{i+1} - x_i) + (f_i'' - f_j'')(x_{i+1} - x_i) + \omega(x_i, x_{i+1} - x_i),$$

$$i = 1, \dots, n, \quad 1 \leq j \leq n. \quad (3.3)$$

If matrix $f''(x)$ is continuous and nonsingular, then by virtue of the assumed closeness of points x_i the sum of the two last terms of the right-hand side of each equation in system (3.3) must be considerably less than the first term, i.e.

$$A(x_{i+1} - x_i) \approx f''(x_j)(x_{i+1} - x_i), \quad i = 1, \dots, n$$

and this in general means that matrices A and f_j'' , $j = 1, \dots, n$, must be close to each other. It can be easily imagined how the above considerations can be used for the construction of iterative processes of minimization. If $\{x_k\}$ is an arbitrarily constructed sequence which converges to the minimum point of $f(x)$, then in a sufficiently small neighbourhood of the minimum point points $x_k, x_{k-1}, \dots, x_{k-n}$ are close to one another. Therefore having defined matrix A_k by the system of equations

$$A_k(x_{k-i} - x_{k-i-1}) = f'(x_{k-i}) - f'(x_{k-i-1}), \quad i = 0, 1, \dots, n-1$$

we can construct the $(k+1)$ -th approximation using now the formula

$$x_{k+1} = x_k - \alpha_k A_k^{-1} f'_k, \quad \alpha_k > 0. \quad (3.4)$$

If matrix A_k happens to be sufficiently close to matrix f_k'' , then direction $p_k = -A_k^{-1} f'_k$ will be close to direction $-(f_k'')^{-1} f'_k$ (i.e. to the direction of motion in Newton's method) and therefore will be the direction of descent. If we continue in a similar way to determine matrices $A_{k+1}, A_{k+2}, \dots$, by virtue of their being close to matrices $f_{k+1}'', f_{k+2}'', \dots$ process (3.4) must be close as to its properties to Newton's method. At the same time method (3.4) does not require the calculation of the second derivatives of the function.

On the basis of the preceding reasoning it proves possible to construct a whole class of descent processes with a superlinear rate of convergence and whose implementation does not require the calculation of second derivatives of the function. We call these processes methods of dual directions. The origin of this name will become clear later on when we shall discuss methods for calculating matrix A_k^{-1} and vector p_k .

We now turn to the strict substantiation of methods of the (3.4) type.

Substantiation of the Methods

Suppose that $f(x)$ is a function which has continuous first and second derivatives. Given an infinite sequence of elements $\{x_k\}$. We take the sequence $\{y_k\}$ corresponding to $\{x_k\}$ in accordance with the formula

$$y_k = x_k + r_k \quad (3.5)$$

where vectors r_k are such that the following two conditions are satisfied:

(1) If Δ_k is a determinant whose columns are vectors $\frac{r_k}{\|r_k\|}, \dots, \frac{r_{k-n+1}}{\|r_{k-n+1}\|}$, then with any $k \geq n-1$ we have $|\Delta_k| \geq \varepsilon$, ε is an arbitrarily small positive number.

(2) $\|r_k\| \rightarrow 0$ as $k \rightarrow \infty$. In other respects the choice of vectors r_k is arbitrary.

The first of the requirements that vectors r_k must satisfy is, in point of fact, the requirement of their linear independence.

Lemma 3.1. *Let $\{x_k\}$ be a bounded sequence, where $\|x_{k+1} - x_k\| \rightarrow 0$ as $k \rightarrow \infty$ and with any $k \geq n-1$ matrix A_k be defined by the following system of equations*

$$A_k r_{k-i} = e_{k-i}, \quad i = 0, 1, \dots, n-1 \quad (3.6)$$

where $e_{k-i} = f'(y_{k-i}) - f'(x_{k-i})$ and r_k, y_k are elements of sequence (3.5). Then we have

$$\lim_{k \rightarrow \infty} \|A_k - f''(x_k)\| = 0.$$

Proof. Using formula (5.1) of Chap. I to give the operator $f'(x)$ in terms of its derivatives, we can write

$$\begin{aligned} f'(y_{k-i}) - f'(x_{k-i}) &= \int_0^1 f''(x_{k-i} + \tau(y_{k-i} - x_{k-i})) r_{k-i} d\tau \\ &= \int_0^1 f''(x_{k-i}) r_{k-i} d\tau + \int_0^1 [f''(x_{k-i} + \tau r_{k-i}) - f''(x_{k-i})] r_{k-i} d\tau \\ &= f''(x_{k-i}) r_{k-i} + \int_0^1 [f''(x_{k-i} + \tau r_{k-i}) - f''(x_{k-i})] r_{k-i} d\tau. \end{aligned}$$

Using this expression we have

$$\begin{aligned} (A_k - f''(x_k)) r_{k-i} &= (f''(x_{k-i}) - f''(x_k)) r_{k-i} \\ &+ \int_0^1 [f''(x_{k-i} + \tau r_{k-i}) - f''(x_{k-i})] r_{k-i} d\tau. \end{aligned}$$

Introducing the notation $A_k - f_k'' = B_k$ we obtain

$$\begin{aligned} & \|B_k r_{k-i}\| \leq \|f_{k-i}'' - f_k''\| \|r_{k-i}\| \\ & + \sup_{0 \leq \tau \leq 1} \|f''(x_{k-i} + \tau r_{k-i}) - f''(x_{k-i})\| \|r_{k-i}\|. \end{aligned} \quad (3.7)$$

Since $\{x_k\}$ is a bounded sequence, with any k we have $x_k \in Q$, $Q \subset E^n$ is a closed bounded set. Function $f''(x)$ is uniformly continuous in set Q ; consequently $\|f_{k-i}'' - f_k''\| = \lambda_{k-i} \rightarrow 0$ and

$$\sup_{0 \leq \tau \leq 1} \|f''(x_{k-i} + \tau r_{k-i}) - f''(x_{k-i})\| = \mu_{k-i} \rightarrow 0 \text{ as } k \rightarrow \infty.$$

Thus it follows from (3.7) that

$$\|B_k r_{k-i}\| \leq (\lambda_{k-i} + \mu_{k-i}) \|r_{k-i}\| = h_{k-i} \|r_{k-i}\| \quad (3.8)$$

where $h_{k-i} \rightarrow 0$ as $k \rightarrow \infty$.

According to the definition of the operator norm, $\|B_k\| = \max_{\|z\|=1} \|B_k z\|$. Let the maximum be attained at element z_k . If

$$z_k = \delta_k \frac{r_k}{\|r_k\|} + \dots + \delta_{k-n+1} \frac{r_{k-n+1}}{\|r_{k-n+1}\|},$$

then because of the condition $|\Delta_k| \geq \varepsilon > 0$, the coefficients δ_{k-i} will be bounded $|\delta_{k-i}| \leq C$, $i = 0, 1, \dots, n-1$. Using the expression of vector z_k we obtain

$$\|B_k\| = \|B_k z_k\| = \left\| \sum_{i=0}^{n-1} \delta_{k-i} B_k \frac{r_{k-i}}{\|r_{k-i}\|} \right\| \leq \sum_{i=0}^{n-1} \left\| \delta_{k-i} B_k \frac{r_{k-i}}{\|r_{k-i}\|} \right\|.$$

Hence taking into account (3.8) and the fact that $|\delta_{k-i}|$ is bounded we have

$$\|B_k\| \leq \sum_{i=0}^{n-1} |\delta_{k-i}| \frac{h_{k-i} \|r_{k-i}\|}{\|r_{k-i}\|} = \sum_{i=0}^{n-1} |\delta_{k-i}| h_{k-i} \rightarrow 0$$

as $k \rightarrow \infty$. The lemma is proved.

The results of the above lemma open the way to the construction of methods of the type (3.4).

Lemma 3.2. *If $f(x)$ is a continuously differentiable strongly convex function and sequence $\{x_k\}$ is such that $f_{k+1} \leq f_k$ and $(f'_k, x_{k+1} - x_k) \rightarrow 0$ as $k \rightarrow \infty$, then $\|x_{k+1} - x_k\| \rightarrow 0$.*

Proof. According to condition $f_{k+1} \leq f_k$ we have $x_{k+1} \in S_k$, $S_k = \{x: f(x) \leq f(x_k)\}$ with any k . Set S_k is strongly convex since $f(x)$

is a strongly convex function (lemma 2.8 of Chap. I). Then there is a positive number $\lambda > 0$ such that any point $\frac{x_{k+1} + x_k}{2} + \xi$, where $\|\xi\| \leq \lambda \|x_{k+1} - x_k\|^2$, is an internal point of set S_k . Let $\frac{x_{k+1} - x_k}{2} = v + \omega$ where $v \in T_k$, T_k is a plane tangent to set S_k at point x_k and $\omega \perp T_k$. Then noting that $f'(x_k) \perp T_k$ we obtain

$$\frac{1}{2} |(f'_k, x_{k+1} - x_k)| = |(f'_k, v + \omega)| = \|f'_k\| \|\omega\|.$$

But $\|\omega\| > \|\xi\|$ since otherwise in addition to point x_k set S_k and plane T_k would have other points in common, which contradicts the strong convexity of S_k . Therefore

$$\frac{1}{2} |(f'_k, x_{k+1} - x_k)| \geq \lambda \|f'_k\| \|x_{k+1} - x_k\|^2.$$

Hence if $\|f'_k\| \not\rightarrow 0$ then $\|x_{k+1} - x_k\| \rightarrow 0$. But if $\|f'_k\| \rightarrow 0$ then, since $f(x)$ is strongly convex, the maximum diameter of set S_k $d_k \rightarrow 0$, which implies that $\|x_{k+1} - x_k\| \rightarrow 0$. The lemma is proved.

We can now discuss the properties of process (3.4). We shall study this process assuming that the value of parameter α_k is chosen according to condition (2.2) taking into account that in this case $p_k = -A_k^{-1}f'_k$. The function being minimized is assumed to be smooth and strongly convex. We motivate the possibility of choosing parameter α_k in the same way as was done in the preceding sections. Expanding the function in Taylor's series, we obtain the following estimate:

$$f_{k+1} - f_k \leq \alpha_k (f'_k, p_k) \left[1 + \frac{\alpha_k}{2} \frac{(A_k p_k, p_k)}{(f'_k, p_k)} + \frac{\alpha_k a_k}{2} \frac{\|p_k\|^2}{(f'_k, p_k)} \right]$$

where $a_k = \|f'_{k_c} - A_k\|$, $x_{k_c} = x_k + \theta(x_{k+1} - x_k)$, $\theta \in [0, 1]$. Note that

$$(f'_k, p_k) = -(A_k p_k, p_k) = -(f'_k, A_k^{-1} f'_k). \quad (3.9)$$

Taking (3.9) into account we have

$$f_{k+1} - f_k \leq \alpha_k (f'_k, p_k) \left[1 - \frac{\alpha_k}{2} + \frac{\alpha_k a_k}{2} \frac{\|p_k\|^2}{(f'_k, p_k)} \right].$$

Hence condition (2.2) will be always satisfied if α_k satisfies the following inequality:

$$1 - \frac{\alpha_k}{2} + \frac{\alpha_k a_k}{2} \frac{\|p_k\|^2}{(f'_k, p_k)} \geq \varepsilon. \quad (3.10)$$

Theorem 3.1. *If $f(x)$ is a twice continuously differentiable function for which conditions (2.4) are valid, matrix A_k with any $k \geq n - 1$*

is defined by system (3.6) and satisfies the condition

$$(A_k^{-1}f'_k, f'_k) > 0 \quad (3.11)$$

and α_k is determined according to condition (2.2), then whatever the initial point x_0 the following statements are valid for sequence (3.4): $f_{k+1} < f_k$ and $\|x_k - x_*\| \rightarrow 0$ at a superlinear rate of convergence:

$$\|x_{N+l} - x_*\| \leq C\lambda_N \dots \lambda_{N+l} \quad (3.12)$$

where $C, N < \infty$, $\lambda_{N+l} < 1$ with any $l \geq 0$, $\lambda_i \rightarrow 0$ as $i \rightarrow \infty$.

Proof. In order to make use of the results of lemma 3.1 we must first of all show that the conditions of the theorem imply that $\|x_{k+1} - x_k\| \rightarrow 0$ for sequence (3.4).

According to conditions (3.9) and (3.11) we have $(f'_k, p_k) < 0$ with any k . Hence it follows, first, that there is always a value $\alpha_k \neq 0$ such that inequality (3.10) is satisfied (and consequently also (2.2)), secondly, by (2.2) we shall have $f_{k+1} < f_k$. This means that $x_{k+1} \in S = \{x: f(x) \leq f(x_0)\}$ with any k and besides since $f(x)$ has a lower bound, $f_k - f_{k+1} \rightarrow 0$; by virtue of this it follows from (2.2) that

$$\alpha_k (f'_k, p_k) = (f'_k, x_{k+1} - x_k) \rightarrow 0. \quad (3.13)$$

Since $f_{k+1} \leq f_k$ and condition (3.13) is fulfilled, sequence $\{x_k\}$ satisfies the requirements of lemma 3.2. Hence, as $k \rightarrow \infty$

$$\|x_{k+1} - x_k\| \rightarrow 0. \quad (3.14)$$

Thus the conditions of the theorem provide for the fulfillment of all the requirements of lemma 3.1. It follows that

$$\|A_k - f''_k\| \rightarrow 0 \quad (3.15)$$

as $k \rightarrow \infty$. Taking into account conditions (2.4), for any M_1 and m_1 such that $M_1 \geq M$ and $0 < m_1 \leq m$ there is a number L such that with $k \geq L$ for any $y \in E^n$ we have

$$m_1 \|y\|^2 \leq (A_k y, y) \leq M_1 \|y\|^2. \quad (3.16)$$

Because of conditions (3.9) and (3.16) from a certain k on we shall have $(f'_k, p_k) \leq -m_1 \|p_k\|^2$. Consequently inequality (3.10) will always be satisfied if the following inequality is satisfied:

$$1 - \frac{\alpha_k}{2} - \frac{\alpha_k a_k}{2} \frac{1}{m_1} \geq \varepsilon, \quad 0 < \varepsilon < \frac{1}{2}. \quad (3.17)$$

Noting that $a_k \leq M + M_1 < \infty$ it is easily ascertained that there is a constant $\alpha > 0$ such that with any k inequality (3.17) will be satisfied with $\alpha_k \geq \alpha$. By virtue of this it follows from (3.14) that $\|p_k\| = \frac{1}{\alpha_k} \|x_{k+1} - x_k\| \rightarrow 0$.

Hence

$$\|f'_k\| = \|A_k p_k\| \leq M_1 \|p_k\| \rightarrow 0.$$

The last condition as shown by inequality (1.12) means that $x_k \rightarrow x_*$. Let us establish that estimate (3.12) holds.

Since $\|p_k\| \rightarrow 0$ and condition (3.15) is fulfilled and the second derivatives of function $f(x)$ are uniformly continuous on set S , we have as $k \rightarrow \infty$

$$\alpha_k \leq \|f''(x_k + \theta(x_{k+1} - x_k)) - f''(x_k)\| + \|f''(x_k) - A_k\| \rightarrow 0$$

and it follows that with any $0 < \varepsilon < \frac{1}{2}$ from a certain iteration on inequality (3.17) will be satisfied with $\alpha_k = 1$. This means that process (3.4) will be implemented with a step equal to unity.

The left-hand side in equality of estimates (3.16) means that with $k \geq L$ we shall have $\|A_k^{-1}\| \leq \frac{1}{m_1}$ (lemma 2.9 of Chap. I). Together with the possibility of inverting matrix A_k with any $k \geq n - 1$ this estimate makes it possible to conclude that there is a constant M_2 such that $\|A_k^{-1}\| \leq M_2$ with $k \geq n - 1$. Taking this into account we can establish the bounds on the rate of convergence like we did in theorem 2.1:

$$\begin{aligned} \|x_{k+1} - x_*\| &\leq \|I - A_k^{-1} f''_{kc}\| \|x_k - x_*\| \\ &\leq \|A_k^{-1}\| \|A_k - f''_{kc}\| \|x_k - x_*\| \leq M_2 \|A_k - f''_{kc}\| \|x_k - x_*\| \end{aligned}$$

or $\|x_{k+1} - x_*\| \leq \lambda_k \|x_k - x_*\|$ where $\lambda_k = M_2 \|A_k - f''_{kc}\|$. Since $\|A_k - f''_{kc}\| \leq \|A_k - f''_k\| + \|f''_k - f''(x_k + \theta(x_k - x_*))\| \rightarrow 0$, there is a number N such that with $k = N + l$, $l = 0, 1, \dots$ we shall have $\lambda_{N+l} < 1$ and as $l \rightarrow \infty$, $\lambda_{N+l} \rightarrow 0$.

Setting $\|x_N - x_*\| = C$ and taking into account our remarks about the values of λ_k , we obtain estimate (3.12). The theorem is proved.

Condition (3.11) used in the theorem, because of (3.9), means that the direction p_k is the direction of descent of $f(x)$. It can occur that $(A_k^{-1} f'_k, f'_k) \leq 0$ at some iterations of the process. In this case we can either change vector r_k and construct a new matrix A_k^{-1} (such that condition (3.11) be satisfied) or make a step in the direction of the antigradient. The number of such steps will always be finite since in descending along the antigradient $\|x_{k+1} - x_k\| \rightarrow 0$ and if vectors r_k satisfy the requirements formulated above, then $\|A_k - f''_k\| \rightarrow 0$. Consequently according to conditions (3.16) and (3.9) from a certain iteration on we shall have of necessity $(A_k^{-1} f'_k, f'_k) > 0$. However if with a certain k the condition $(A_k^{-1} f'_k, f'_k) = -(p_k, f'_k) < 0$ is fulfilled, then it is easier to change in formula (3.4) the sign of the scalar factor α_k ; then the motion from point x_k will be in the direction $-p_k$, i.e. in the direction of descent.

Construction of Various Algorithms

The requirements which must be satisfied by vectors r_k used in constructing sequence (3.5) are not strict and leave us a great freedom of choice of these vectors. This makes it possible to construct different algorithms of type (3.4) since different sequences $\{r_k\}$ will define (by (3.6)) different sequences of matrices A_k .

Let us discuss some possible ways of constructing vectors r_k . We can take as r_k vectors directed along the axes of coordinates. For example, if $r_0 = \lambda_0 v_1$, then with $k = tn + i$, where t is an integer and $i = 0, 1, \dots, n-1$, we have $r_k = \lambda_k v_{i+1}$; v_{i+1} is the unit vector of the corresponding axis and λ_k is a numerical factor such that $\lambda_k \rightarrow 0$ as $k \rightarrow \infty$. Such a choice of vector r_k guarantees that the condition $|\Delta_k| \geq \varepsilon$ will be satisfied. In this case in order to determine matrix A_k , it is necessary to calculate at each iteration the derivatives at two points, x_k and y_k . The law of the decrease of λ_k may be chosen arbitrarily; computational practice shows however that the maximum rate of convergence is obtained with monotonically diminishing λ_k ; one can, for instance, assume that $\lambda_k = \frac{1}{k^2}$.

Another possible method of determining vectors r_k is as follows. With $k \geq n-1$ we can, instead of (3.5), use directly sequence (3.4), i.e. assume $r_k = x_{k+1} - x_k = -\alpha_k A_k^{-1} f'_k$. In fact, the proof of theorem 3.1 shows clearly that if A_k is an arbitrary matrix which satisfies only condition (3.11) and α_k is chosen according to condition (2.2), then $\|x_{k+1} - x_k\| \rightarrow 0$ as $k \rightarrow \infty$. Consequently, if we use sequence (3.4) for constructing vectors r_k , then the requirement $\|r_k\| \rightarrow 0$ will of necessity be fulfilled and we only need to establish $|\Delta_k| \geq \varepsilon$. If this condition is not satisfied with a certain k , another vector r_k must be chosen (using not (3.4) but a new formula). In such an algorithm, for determining matrix A_k at every iteration (where sequence (3.4) provides the fulfillment of the requirements to be met by vectors r_k) the gradient must be calculated only at one point x_k .

Of course, other methods of constructing vectors r_k may be used.

In the system of equations (3.6) which defines matrix A_k only one new vector r_k and the corresponding vector e_k are used with any k ; the remaining vectors $r_{k-1}, \dots, r_{k-n+1}$ and $e_{k-1}, \dots, e_{k-n+1}$ are constructed from preceding iterations. The system (3.6) can be modified so that at each iteration of process (3.4) an arbitrary number of vectors $r_{k-i_1}, \dots, r_{k-i_j}$, $1 \leq j \leq n$ (and of their corresponding vectors $e_{k-i_1}, \dots, e_{k-i_j}$) be renewed and the remaining $n-j$ vectors $r_{k-1-j+1}, \dots, r_{k-i_n}$ be taken from preceding iterations. In this case system (3.6) should preferably be written in the form

$$A_k r_i = e_i, \quad i = 1, \dots, n. \quad (3.18)$$

If the requirements of lemma 3.1 are retained, then repeating its proof word for word, it can be ascertained that for matrix A_k defined by system (3.18) the condition $\|A_k - f''_k\| \rightarrow 0$ as $k \rightarrow \infty$ is also satisfied.

Using different methods for constructing vectors r_i in system (3.18) one can obtain several well known minimization algorithms. For instance, if we set $r_i = v_{ki}$ (and $y_i = x_k + v_{ki}$) where v_{ki} is a vector directed along the i -th axis of coordinates and such that $\|v_{ki}\| \rightarrow 0$ as $k \rightarrow \infty$, then system (3.18) will take the following form:

$$A_k v_{ki} = f'(x_k + v_{ki}) - f'(x_k), \quad i = 1, \dots, n.$$

Matrix A_k defined by this system is a finite differences analogue of the matrix of second derivatives $f''(x_k)$; thus in this case process (3.4) transforms into a finite differences analogue of Newton's method with adjustment of step length. On the basis of theorem 3.1 it can be asserted that Newton's finite differences method with adjustment of step length converges from any initial approximation at a super-linear rate. Assuming that matrix $f''(x)$ satisfies Lipschitz' condition (2.8), it can be shown, using the preceding results, that if $\|v_{ki}\| \leq \|f'_k\|$, Newton's finite differences method converges at a quadratic rate. This can be seen from the following argument. In lemma 3.1 the bound on the quantity $B_k r_i = B_k v_{ki}$ takes the form

$$\|B_k v_{ki}\| \leq \sup_{0 \leq \tau \leq 1} \|f''(x_k + \tau v_{ki}) - f''(x_k)\| \|v_{ki}\|.$$

Hence, using (2.8), $\|B_k v_{ki}\| \leq R \|v_{ki}\|^2$. From this inequality and the estimate $\|v_{ki}\| \leq \|f'_k\|$, as in lemma 3.1, we obtain

$$\|A_k - f''_k\| = \|B_k\| \leq R \sum_{i=1}^n |\delta_i| \|v_{ki}\| \leq R \sum_{i=1}^n |\delta_i| \|f'_k\| = \tilde{R} \|f'_k\|.$$

According to (2.4) the gradient $f'(x)$ satisfies Lipschitz' condition with constant M . Consequently,

$$\|B_k\| \leq \tilde{R} \|f'_k\| = \tilde{R} \|f'_k - f'_*\| \leq \tilde{R} M \|x_k - x_*\|.$$

The bounds on the rate of convergence obtained in theorem 3.1 can now be defined more exactly as follows:

$$\begin{aligned} \|x_{k+1} - x_*\| &\leq M_2 \|A_k - f''_{kc}\| \|x_k - x_*\| \\ &\leq M_2 (\|A_k - f''_k\| + \|f''_k - f''(x_k + \theta(x_k - x_*))\|) \|x_k - x_*\| \\ &\leq M_2 (\tilde{R} M + R) \|x_k - x_*\|^2, \end{aligned}$$

i.e.

$$\|x_{k+1} - x_*\| \leq C \|x_k - x_*\|^2.$$

Let us describe another method of choosing vectors r_i . We set $r_i = -\lambda_k f'(y_i)$, $y_i = x_k$, $y_{i+1} = y_i + r_i$, $i = 1, 2, \dots, n$. Then the system (3.18) takes the following form:

$$-\lambda_k A_k f'(y_i) = f'(y_{i+1}) - f'(y_i), \quad i = 1, \dots, n.$$

Matrix A_k defined by such a system of equations (accurate to a numerical factor λ_k) is used to construct an iterative process (Aitken and Steffensen's method). We shall not study the properties of this method here. Note only that by varying the value of factor λ_k a quadratic rate of convergence of the process can be obtained.

Determining Vector p_k

The amount of work required to compute vector p_k determines to a considerable extent the computational effort in process (3.4).

We shall now consider a method of constructing vector $p_k = -A_k^{-1} f'_k$ which makes use of the specific properties of system (3.6) that defines matrix A_k . By taking account of these properties one can considerably simplify the construction of vector p_k ; we begin with the inversion of matrix A_k .

The necessary condition for the existence of matrix A_k^{-1} is that matrix A_k be nonsingular and this in turn necessitates linear independence of the vector system $e_k, \dots, e_{k-n+1}$. With sufficiently large k , matrix A_k is nonsingular as shown by (3.16). However, at some iterations of the initial stage of process (3.4) the vector system $e_k, \dots, e_{k-n+1}$ can prove linearly dependent. In this case we can either change one of the vectors r_{k-i} or make a step in the direction of the antigradient this causing a change of system $e_k, \dots, e_{k-n+1}$. We assume in what follows that with any $k \geq n - 1$ system $e_k, \dots, e_{k-n+1}$ is linearly independent.

In this case system (3.6) can be written in the form

$$A_k^{-1} e_{k-i} = r_{k-i}, \quad i = 0, 1, \dots, n - 1$$

or in the form of a matrix equation:

$$A_k^{-1} E_k = R_k \quad (3.19)$$

where E_k, R_k are matrices whose columns are coordinates of vectors e_{k-i} and r_{k-i} respectively. From (3.19) we obtain

$$A_k^{-1} = R_k E_k^{-1}. \quad (3.20)$$

Thus in order to construct matrix A_k^{-1} it is necessary first of all to calculate matrix E_k^{-1} . It is known from linear algebra (see, for instance, D. K. Faddeev and V. N. Faddeeva) that rows of matrix E_k^{-1} will be vectors whose basis $s_k, \dots, s_{k-n+1}$ is dual for or biorthogonal to basis $e_k, \dots, e_{k-n+1}$. Recall that linearly independent systems

of vectors $a_1, \dots, a_n$ and $b_1, \dots, b_n$ are called *dual* if they satisfy the conditions

$$(a_i, b_j) = 0 \text{ with } i \neq j, \quad (a_j, b_i) = 1.$$

If $s_k, \dots, s_{k-n+1}$ is the dual of basis $e_k, \dots, e_{k-n+1}$, then (according to the relations of duality) $S_k^* E_k = I$, where S_k is a matrix whose columns are vectors s_{k-i} . It follows that $S_k^* = E_k^{-1}$.

Each of the matrices E_k , $k = 0, 1, \dots$ differs from the next one on the left and on the right side only by one column. By virtue of this fact the process of constructing the basis $s_k, \dots, s_{k-n+1}$ can be performed by recursive relations and this to a considerable extent reduces computational efforts.

Suppose we have calculated matrix E_k^{-1} , i.e. have constructed the basis $s_k, \dots, s_{k-n+1}$. Let us construct the system of vectors $\bar{s}_{k+1}, \bar{s}_k, \dots, \bar{s}_{k-n+2}$ as follows:

$$\bar{s}_{k+1} = \frac{s_{k-n+1}}{(s_{k-n+1}, e_{k+1})}.$$

$$\bar{s}_{k+1-j} = s_{k+1-j} - (s_{k+1-j}, e_{k+1}) \bar{s}_{k+1}, \quad j = 1, \dots, n-1. \quad (3.21)$$

It follows from the linear independence of vectors $e_{k+1}, e_k, \dots, e_{k-n+2}$ that

$$(s_{k-n+1}, e_{k+1}) \neq 0 \quad (3.22)$$

Indeed, by virtue of the duality of the bases $s_k, \dots, s_{k-n+1}$ and $e_k, \dots, e_{k-n+1}$ we have $(s_{k-n+1}, e_{k-j}) = 0$ with $j = 0, 1, \dots, n-2$ and if we had $(s_{k-n+1}, e_{k+1}) = 0$, then it would follow that vectors $e_{k+1}, \dots, e_{k-n+2}$ are linearly dependent. Thus, *in order to check the linear independence of vectors $e_{k+1}, e_k, \dots, e_{k-n+2}$, it is sufficient to check condition (3.22).*

Let us show that the vector system (3.21) is a basis which is dual for the basis $e_{k+1}, \dots, e_{k-n+2}$.

Indeed,

$$(\bar{s}_{k+1}, e_{k+1}) = \frac{(s_{k+1-n}, e_{k+1})}{(s_{k+1-n}, e_{k+1})} = 1,$$

$$(\bar{s}_{k+1-j}, e_{k+1}) = (s_{k+1-j}, e_{k+1}) - (s_{k+1-j}, e_{k+1}) \frac{(s_{k+1-n}, e_{k+1})}{(s_{k+1-n}, e_{k+1})} = 0,$$

$$(\bar{s}_{k+1}, e_{k+1-j}) = \frac{(s_{k+1-n}, e_{k+1-j})}{(s_{k+1-n}, e_{k+1})} = 0$$

by virtue of the duality of the bases $e_k, \dots, e_{k-n+1}$ and $s_k, \dots, s_{k-n+1}$,

$$(\bar{s}_{k+1-j}, e_{k+1-m}) = (s_{k+1-j}, e_{k+1-m}) - (s_{k+1-j}, e_{k+1}) \frac{(s_{k+1-n}, e_{k+1-m})}{(s_{k+1-n}, e_{k+1})} = \delta_{jm}$$

$j, m = 1, 2, \dots, n - 1$, δ_{jm} is Kronecker's delta ($\delta_{jj} = 1$, $\delta_{jm} = 0$ with $j \neq m$). Consequently, the conditions of duality are satisfied for vector systems $e_{k+1}, \dots, e_{k-n+2}$ and $\bar{s}_{k+1}, \dots, \bar{s}_{k-n+2}$, and this corroborates our statement.

Thus, using recursive relations (3.21), the construction of basis $\bar{s}_{k+1}, \dots, \bar{s}_{k-n+2}$ (i.e. of matrix E_{k+1}^{-1}) is performed quite easily.

We can now derive a simple formula for the determination of the direction of motion p_k . For this purpose let us write equation (3.20) in the following form:

$$A_k^{-1} = \sum_{i=0}^{n-1} r_{k-i} s_{k-i}^* \quad (3.23)$$

(where superscript * denotes a vector-row). Using (3.23) we obtain

$$p_k = -A_k^{-1} f'_k = -\sum_{i=0}^{n-1} r_{k-i} s_{k-i}^* f'_k = -\sum_{i=0}^{n-1} (s_{k-i}, f'_k) r_{k-i}. \quad (3.24)$$

It was this formula for the determination of the direction of descent in methods of the type (3.4) that caused such algorithms to be called methods of dual directions. Using (3.24), formula (3.4) can be written

$$x_{k+1} = x_k - \alpha_k \sum_{i=0}^{n-1} (s_{k-i}, f'_k) r_{k-i} \quad (3.25)$$

or in coordinate form

$$x_{k+1}^v = x_k^v - \alpha_k \sum_{i=0}^{n-1} \sum_{j=1}^n s_{k-i}^j \frac{\partial f(x_k)}{\partial x^j} r_{k-i}^v, \quad v = 1, \dots, n.$$

In practice, sequences of approximations should be always constructed by this formula.

Note also that using expression (3.23) a recursive formula can be obtained for the calculation of matrix A_k^{-1} . We give it here without its derivation:

$$A_{k+1}^{-1} = A_k^{-1} + (r_{k+1} - A_k^{-1} e_{k+1}) \bar{s}_{k+1}^*.$$

It is easy to check its validity by direct multiplication of a matrix constructed by this formula and vectors $e_{k+1}, e_k, \dots, e_{k-n+2}$. It can be seen then that

$$A_{k+1}^{-1} e_{k+1-i} = r_{k+1-i}, \quad i = 0, 1, \dots, n - 1,$$

i.e. that matrix A_{k+1}^{-1} satisfies system (3.6).

The Initial Stage of the Process

Until now we considered the iterative process (3.4) beginning with $k = n - 1$ since for the definition of matrix A_k , n vectors r_k and corresponding vectors e_k are required.

The first iterations of the process ($k < n - 1$) can be performed in various ways. For instance, use can be made of the method of steepest descent: $x_{k+1} = x_k - \alpha_k f'_k$, $\alpha_k > 0$, $k = 0, 1, \dots, n-2$. In order to secure uniformity of the algorithm from the first iteration on, we can proceed as follows with $0 \leq k < n - 1$.

Set $A_0^{-1} = I$. Present this matrix in the form $A_0^{-1} = R_0 E_0^{-1}$, where $R_0 = I$, $E_0^{-1} = I$, or, using (3.23),

$$A_{0\text{m}}^{-1} = \sum_{i=0}^{n-1} r_{0-i} s_{0-i}^*$$

where $r_0, r_{-1}, \dots, r_{-n+1}$ and $s_0, s_{-1}, \dots, s_{-n+1}$ are vectors of a unity orthonormalized basis. Taking this into account, we have:

$$x_1 = x_0 - \alpha_0 \sum_{i=0}^{n-1} (f'_0, s_{0-i}) r_{0-i}.$$

Further, having calculated vectors r_1 and e_1 by (3.21) we construct the basis:

$$\bar{s}_1 = \frac{s_{-n+1}}{(s_{-n+1}, e_1)}, \quad \bar{s}_{1-j} = s_{1-j}, \quad - (s_{1-j}, e_1) s_1, \quad j = 1, \dots, n-1$$

and the next approximation:

$$x_2 = x_1 - \alpha_1 \sum_{i=0}^{n-1} (f'_1, \bar{s}_{1-i}) r_{1-i}.$$

The construction of the successive iterations is straightforward.

Minimization of Quadratic Form

Let us consider as an example the application of the methods of dual directions to the finding of the minimum point of a quadratic function. Let

$$f(x) = \frac{1}{2} (Ax, x) + (b, x) + c$$

where A is an $n \times n$ symmetric, strictly positive definite matrix with constant elements: $(Ax, x) > 0$ for any $x \neq 0$, b is a vector, c is a scalar quantity. The gradient of this function is $f'(x) = Ax + b$ the vector

$$e_i = f'(x + r_i) - f'(x) = Ar_i. \quad (3.26)$$

Therefore, if $r_1, \dots, r_n$ is a linearly independent system of vectors, $s_1, \dots, s_n$ is the basis dual for the basis $e_1, \dots, e_n$, then because of (3.20) and (3.23), we have

$$A_n^{-1} = R_n E_n^{-1} = \sum_{i=1}^n r_i s_i^*.$$

Now, since it follows from (3.26) that matrix A is defined by the system of equations $Ar_i = e_i$, $i = 1, 2, \dots, n$, we can write

$$A^{-1} = R_n E_n^{-1} = \sum_{i=1}^n r_i s_i^*, \quad (3.27)$$

i.e. $A_n^{-1} = A^{-1}$. Hence

$$x_{n+1} = x_n - A_n^{-1} f'_n = x_n - A^{-1} (Ax_n + b) = -A^{-1}b \quad (3.28)$$

and $f'_{n+1} = -AA^{-1}b + b = 0$, i.e. $x_{n+1} = x_*$.

Thus, in order to minimize a quadratic function by the method of dual directions we have to calculate the gradient of the function at $n + 1$ points and construct a basis which is dual for the basis of vectors $e_1, \dots, e_n$. If we consider the process of successive calculations of vectors $e_1, \dots, e_n$ as a certain iterative procedure, then it can be said that methods of dual directions make it possible to minimize a quadratic function after a finite number of steps.

Note also that the problem under consideration is equivalent to solving a system of linear equations $Ax = -b$. Consequently, methods of dual directions make it possible to solve a system of linear equations by performing a finite number of iterations.

Discussion of Properties of the Methods

Methods of dual directions make it possible to solve the problem of minimizing a strictly convex smooth function whatever the initial approximation chosen and the rate of convergence of the sequence $\{x_k\}$ to the solution is superlinear. The method of choosing parameter α_k guarantees the determination of the required value of α_k after a finite number of reductions. Of course, in process (3.4) as in the methods described in the preceding sections α_k can be chosen under the condition of obtaining the minimum function value in the direction of motion; however, such a method is more laborious.

The methods of the class under consideration approach Newton's method as to the estimate of their rate of convergence. Let us compare the labour per iteration in the methods of dual directions and in Newton's method.

In processes of type (3.4) with matrix A_k defined by system (3.6) in order to calculate matrix A_k^{-1} we have to calculate vector e_k and

then using recursive formulas (3.21) to construct a basis dual for the basis $e_k, \dots, e_{k-n+1}$.

For constructing vector e_k as stated in the subsection on p. 74 it is necessary to calculate the gradient of the function at one or two points. The amount of work required to construct the dual basis by formulas (3.21) is only $1/n$ of that required by the usual methods (D. K. Faddeev and V. N. Faddeeva) (it is reduced even more if we check conditions that provide the inequality to zero of the denominator in the general formulas for the construction of a dual basis).

Thus first, the methods of dual directions, as distinct from Newton's method, do not require calculation of the second derivatives of the function. We compare the dual direction methods to Newton's finite differences method. We find that the amount of computations in the former methods required for the construction of matrix A_k^{-1} is about $\frac{1}{n}$ as much, since Newton's method necessi-

tates the calculation at each iteration of derivatives at $n + 1$ points and the inversion of matrix A_k without using recursive relations.

In Newton's finite differences method, in order to determine the direction of motion p_k , instead of inverting the matrix one can solve a system of linear equations (like it can be done in the usual Newton method). In such a case the quantitative estimation of the computational effort in methods of the type (3.4) and in Newton's method depends on the method of solving the system of equations; however, the ratio is also approximately equal to n . For instance, if for solving the system of linear equations we use the method of dual directions (see the subsection on p. 79), we have, in practice, to calculate matrix A_n^{-1} without using recursive relations, which, as was mentioned above, requires an n times greater amount of calculations as compared to that required by formulas (3.21).

Computational effort is nearly the same in solving a system of linear equations by methods of conjugate directions; we shall discuss this method in the following section.

Thus methods of dual directions converging at a rate close to that of Newton's method require at the same time a far lesser amount of calculations per iteration. The shortcoming of these methods is that their implementation on computers requires a larger storage capacity of the computer, since it is necessary to memorize two systems of vectors $r_k, r_{k-1}, \dots, r_{k-n+1}$ and $s_k, s_{k-1}, \dots, s_{k-n+1}$, i.e. actually two $n \times n$ matrices. This is an obstacle to the application of methods of dual directions to the solving of large size problems on computers with a limited working storage. It should be noted, however, that this shortcoming is partially compensated if we use vectors directed along coordinate axes as vectors r_k , for in this case we have to store in the computer memory only one n -dimensional vector instead of the system $r_k, \dots, r_{k-n+1}$.

4. METHODS OF CONJUGATE DIRECTIONS. MINIMIZATION OF QUADRATIC FUNCTIONS

Conjugate Directions and Their Properties

Let us turn again to the problem of minimizing quadratic functions of the form

$$f(x) = \frac{1}{2}(Ax, x) + (b, x) + c \quad (4.1)$$

where $(Ax, x) > 0$ with any $x \neq 0$; we considered this problem in the preceding section (the subsection on p. 79). It is easy to ascertain that the problem of quadratic function minimization can be reduced to the inversion of matrix A ; if matrix A^{-1} is known, then the solution is immediately found by using formulas (3.28):

$$x_* = x_0 - A^{-1}f'_0 = -A^{-1}b \quad (4.2)$$

where x_0 is an arbitrary point.

If one calculates matrix A^{-1} using expression (3.27), it is necessary, having chosen an arbitrary linearly independent system of vectors $p_0, \dots, p_{n-1}$ (we use here the notation p_i instead of r_i), to calculate the corresponding vectors

$$\tilde{e}_i = f'(x_i + p_i) - f'(x_i) = Ap_i, \quad i = 0, 1, \dots, n-1 \quad (4.3)$$

where x_i are arbitrary points, and to construct a basis $s_0, \dots, s_{n-1}$ dual for the basis $\tilde{e}_0, \dots, \tilde{e}_{n-1}$, i.e. which satisfies the conditions

$$(s_i, \tilde{e}_i) = 1, \quad (s_i, \tilde{e}_j) = 0 \text{ with } i \neq j. \quad (4.4)$$

These relations, because of (4.3), can be written in the form

$$(s_i, Ap_i) = 1, \quad (s_i, Ap_j) = 0, \quad i \neq j. \quad (4.5)$$

Of particular interest is the case in which vectors $p_0, \dots, p_{n-1}$ are *A-orthogonal* or, as they are sometimes called, *conjugate*, i.e. such that

$$(p_i, Ap_j) = 0, \quad i \neq j. \quad (4.6)$$

The system of (nonzero) vectors $p_0, \dots, p_{n-1}$ which satisfies conditions (4.6) is linearly independent (being orthogonal in a metric, defined by a nonsingular matrix) and accordingly can be used to determine vectors e_i by formulas (4.3); vectors s_i which satisfy (4.5) in this case can be calculated by very simple formulas:

$$s_i = \frac{p_i}{(Ap_i, p_i)}, \quad i = 0, 1, \dots, n-1. \quad (4.7)$$

Thus if vectors $p_0, \dots, p_{n-1}$ are A -orthogonal, matrix A^{-1} is calculated by the following formula (see (3.27)):

$$A^{-1} = \sum_{i=0}^{n-1} p_i s_i^* = \sum_{i=0}^{n-1} \frac{p_i p_i^*}{(A p_i, p_i)}, \quad (4.8)$$

i.e. the problem of the inversion of matrix A , and thus that of the minimization of function $f(x)$ is solved quite easily.

Let us now consider the problem of determining point x_* with the aid of conjugate vectors from a somewhat different viewpoint; at the same time we shall study a number of interesting properties of conjugate directions.

Since $p_0, \dots, p_{n-1}$ is the basis of space E^n , point x_* may be presented in the following form:

$$x_* = x_0 + \sum_{i=0}^{n-1} \alpha_i p_i. \quad (4.9)$$

But by (4.2) and (4.8),

$$x_* = x_0 - \sum_{i=0}^{n-1} \frac{p_i p_i^*}{(A p_i, p_i)} f'_0. \quad (4.10)$$

It follows from (4.9) and (4.10) that

$$x_0 + \sum_i \alpha_i p_i = x_0 - \sum \frac{p_i p_i^*}{(A p_i, p_i)} f'_0$$

or in another form

$$x_0 + \sum_i \alpha_i p_i = x_0 - \sum_i \frac{(f'_0, p_i)}{(A p_i, p_i)} p_i. \quad (4.11)$$

Since a vector has only one resolution along the basis axes, the last equality determines the values of coefficients α_i in the expansion (4.9)

$$\alpha_i = -\frac{(f'_0, p_i)}{(A p_i, p_i)} = -\frac{(f'_0, p_i)}{(\tilde{e}_i, p_i)}, \quad i = 0, 1, \dots, n-1. \quad (4.12)$$

Thus if a certain system of conjugate vectors is known, then the minimum point of quadratic function (4.1) is easily found by formulas (4.9), (4.12).

The procedure of determining point x_* by formula (4.9) can be considered as a process of construction of successive points:

$$x_{i+1} = x_i + \alpha_i p_i, \quad i = 0, 1, \dots, n-1 \quad (4.13)$$

where parameters α_i are determined by formulas (4.12). It follows that *using the method of conjugate directions one can solve the problem of quadratic function minimization after performing a finite number of steps not exceeding n* (the number of points in the iterative process (4.13) can prove less than n if some of the coefficients α_i in expansion (4.9) prove equal to zero, i.e. if for some i we have $(f'_0, p_i) = 0$). The above property of the method of conjugate directions is the most important one. It shows how effective is the application of conjugate vectors to quadratic function minimization; this is the reason of wide application of methods of conjugate directions.

As an interesting corollary to the result obtained, it can be shown that *point x_i constructed by formulas (4.13), (4.12) is the minimum point of function (4.1) on the subspace formed by vectors $p_0, \dots, p_{i-1}$ and passing through point x_0* . Let

$$\tilde{x}_i = x_0 + \sum_{k=0}^{i-1} \tilde{\alpha}_k p_k$$

where $\tilde{\alpha}_k$ are arbitrary coefficients. For point $\tilde{x}_i$ to be the minimum of a strictly convex differentiable function in the subspace formed by vectors $p_0, \dots, p_{i-1}$, it is necessary and sufficient (corollary 4.4 of Chap. I) that the following conditions be satisfied:

$$(f'(\tilde{x}_i), p_j) = 0, \quad j = 0, 1, \dots, i-1. \quad (4.14)$$

Now for any $0 \leq j \leq i-1$, we have

$$\begin{aligned} (f'(\tilde{x}_i), p_j) &= (A\tilde{x}_i + b, p_j) = (A(x_0 + \sum_{k=0}^{i-1} \tilde{\alpha}_k p_k) + b, p_j) \\ &= (Ax_0 + b, p_j) + \sum_{k=0}^{i-1} \tilde{\alpha}_k (Ap_k, p_j) = (f'_0, p_j) + \tilde{\alpha}_j (Ap_j, p_j). \end{aligned}$$

Hence taking into account (4.14) we have that point $\tilde{x}_i$ provides the minimum of the function in the subspace, formed by vectors $p_0, \dots, p_{i-1}$ and passing through point x_0 if and only if $(f'_0, p_j) + \tilde{\alpha}_j (Ap_j, p_j) = 0$, i.e.

$$\tilde{\alpha}_j = -\frac{(f'_0, p_j)}{(Ap_j, p_j)}.$$

Now these coefficients coincide with the coefficients α_i calculated by formula (4.12), i.e. point $\tilde{x}_i$ which provides the required minimum coincides with point x_i (4.13). Consequently

$$(f'(x_i), p_j) = 0, \quad j = 0, 1, \dots, i-1. \quad (4.15)$$

It is now clear that minimization of a quadratic function in space E^n by formulas (4.13), (4.12) can be interpreted as a process of

successive minimization of the function in subspaces of $i + 1$ dimensions, $i = 0, 1, \dots, n - 1$, it being necessary to calculate only one coefficient α_i in order to find every time the next minimum point.

Note that in finding α_i by formulas (4.12) we need not practically calculate the matrix of second derivatives A and it is necessary to calculate only vectors $\tilde{e}_i = f'(x_i + p_i) - f'(x_i)$ (see (4.3)), i.e. only the first derivatives of the function.

It is easy to ascertain that formulas (4.12) can be transformed as follows. If x_i is determined by formula (4.13), then

$$\begin{aligned} (f'_0, p_i) &= (f'_0 - f'_1 + f'_1 - \dots - f'_i + f'_i, p_i) \\ &= (-\alpha_0 A p_0 - \alpha_1 A p_1 - \dots - \alpha_{i-1} A p_{i-1} + f'_i, p_i) \end{aligned}$$

and according to the A -orthogonality of vectors $p_0, \dots, p_i$ we have $(f'_0, p_i) = (f'_i, p_i)$.

Consequently

$$\alpha_i = -\frac{(f'_i, p_i)}{(A p_i, p_i)} = -\frac{(f'_i, p_i)}{(\tilde{e}_i, p_i)}, \quad i = 0, 1, \dots, n - 1. \quad (4.16)$$

It follows that if with a certain $0 \leq i \leq n - 1$ in formula (4.13) $\alpha_i = 0$ (i.e. $x_{i+1} = x_i$), then this means that $(f'_i, p_i) = 0$. Combining this equality with (4.15) we obtain

$$(f_{i+1}, p_j) = (f_i, p_j) = 0, \quad j = 0, 1, \dots, i.$$

Thus the fact that the coefficient α_i becomes zero means that the corresponding point x_i provides the minimum of the quadratic function in the subspace formed by vectors $p_0, \dots, p_i$ and passing through point x_0 .

Finally, note that by (4.15) $(f'_i, p_{i-1}) = 0$. This means that the choice of coefficients α_i by formulas (4.12) or (4.16) corresponds to choosing α_i under the condition

$$f(x_i + \alpha_i p_i) = \min_{\alpha} f(x_i + \alpha p_i).$$

Construction of the Methods

In considering in the preceding subsection the effectiveness of methods of conjugate directions for the minimization of a quadratic function, we did not even mention the methods of constructing such vectors and the work involved in this procedure.

We turn to the study of methods of constructing A -orthogonal vectors. Each of these methods determines one or other method of conjugate directions, which consists in the construction of successive approximations to the solution of the problem of minimization of function (4.1) making use of formulas (4.13), (4.12) (or (4.16)).

The effectiveness of the methods of conjugate directions depends directly on the amount of calculations to be performed in order to construct the system of conjugate vectors. If the method chosen for constructing the conjugate vectors proves too laborious, then the effectiveness of the corresponding method of conjugate directions may prove to be low (as compared to algorithms of other classes). Therefore, it is worth while to specify the general requirements which must be satisfied by any method of constructing conjugate vectors in order that the corresponding method of conjugate directions be effective.

First, the process of constructing conjugate vectors should use only calculations of the function and its gradient and should not require the calculation of second derivatives of the function. If this requirement is not satisfied, then minimization of a quadratic function by method (4.13) can involve the need of calculating the matrix of second derivatives, and moreover require the calculation of gradients at several points. Therefore in general, a method of conjugate directions which requires the calculation of the matrix of second derivatives proves less effective than Newton's method (with the possible exception only of problems in which the inversion of matrix A is far more laborious in comparison with its calculation).

Secondly, the information about the function should be used only at points of the sequence (4.13). In other words, the process of constructing conjugate vectors should be such that in determining vector p_i , $0 \leq i \leq n - 1$, the function and its gradient be evaluated only at points $x_0, \dots, x_i$.

It follows from this requirement that one should consider only such methods of constructing conjugate vectors that

$$(f'_i, p_i) = 0, \quad 0 \leq i \leq n - 1 \quad (4.17)$$

is satisfied if and only if $f'_i = 0$. Indeed, if condition (4.17) is satisfied, then by (4.16) $\alpha_i = 0$ and therefore in sequence (4.13) $x_{i+1} = x_i$. This means that at the $(i + 1)$ -th iteration of the process we shall not receive any additional information about the function and therefore shall not be able to construct vector $p_{i+1} \neq p_i$. The process will accordingly degenerate (stop) without reaching the solution if $f'_i \neq 0$.

Thus for any of the methods of constructing conjugate vectors (and for the corresponding method of conjugate directions), the condition

$$(f'_i, p_i) \neq 0 \text{ if } f'_i \neq 0 \quad (4.18)$$

must be satisfied. This condition guarantees that at any of the iterations of the process we shall have $\alpha_i \neq 0$.

In working out algorithms for the construction of conjugate vectors we should assume that condition (4.18) is satisfied. Then, the algorithms having been constructed, it is necessary to check whether this condition is actually satisfied and, if necessary, to impose additional constraints on the algorithm in order to satisfy the condition.

Taking into account the above remarks, let us turn to the actual working out of the relations for the construction of A -orthogonal vectors.

In what follows we use the notations

$$r_i = x_{i+1} - x_i = \alpha_i p_i, \quad e_i = f'_{i+1} - f'_i = \alpha_i A p_i. \quad (4.19)$$

An arbitrary direction of descent of function (4.1) may be chosen to be vector $p_0 = -H_0^* f'_0$, where H_0 is a symmetric, strictly positive definite matrix.

Let us establish the requirements which vector p_k , $1 \leq k \leq n - 1$, must satisfy in order to fulfill the conditions of A -orthogonality:

$$(p_k, A p_j) = 0, \quad 0 \leq j \leq k - 1. \quad (4.20)$$

To this end, we make use of the fact that according to the properties of conjugate directions (see (4.15)) in choosing α_i in process (4.13) by formula (4.16), conditions (4.20) and at the same time also the equality

$$(f'_k, p_j) = 0, \quad 0 \leq j \leq k - 1 \quad (4.21)$$

must be satisfied. If we set

$$p_k = -H_k^* f'_k \quad (4.22)$$

where H_k is an $n \times n$ square matrix, then conditions (4.20) can be written in the following form:

$$(f'_k, H_k A p_j) = 0, \quad 0 \leq j \leq k - 1.$$

Comparison of the equalities obtained with (4.21) shows that if (4.21) is satisfied, then (4.20) will also be satisfied, provided matrix H_k satisfies the relations

$$H_k A p_j = a p_j, \quad 0 \leq j \leq k - 1$$

where a is an arbitrary constant.

Since according to condition (4.18) and the strict convexity of function (4.1) we have $0 < |\alpha_i| < \infty$ with any $0 \leq i \leq n - 1$, equalities (4.20) and (4.21) can be written in the following form:

$$(r_k, e_j) = 0, \quad 0 \leq j \leq k - 1, \quad (4.23)$$

$$(f'_k, r_j) = 0, \quad 0 \leq j \leq k - 1, \quad (4.24)$$

and the conditions for determining matrix H_k can be written as follows:

$$H_k e_j = a r_j, \quad 0 \leq j \leq k - 1. \quad (4.25)$$

Thus the conditions of A -orthogonality (4.20) will be satisfied if matrix H_k which determines vector p_k by formula (4.22) satisfies equations (4.25).

With $k < n - 1$, the number of vector equations (4.25) will be less than n ; it follows that matrix H_k is not uniquely defined. Besides, with different values of constant a the systems of equations for defining matrix H_k will also be different. All this suggests the diversity of algorithms which can be used to construct conjugate directions as we have to use various methods of constructing different matrices H_k .

Since equations (4.25) must be satisfied with any $k = 1, 2, \dots, n - 1$, it is natural to try and construct matrix H_k by recursive relations.

Let us write (4.25) in the following form:

$$(H_{k-1} + \Delta H_{k-1}) e_j = a r_j, \quad 0 \leq j \leq k - 1. \quad (4.26)$$

Since matrix H_{k-1} must satisfy the equations

$$H_{k-1} e_j = a r_j, \quad 0 \leq j \leq k - 2,$$

it follows from (4.26) that matrix ΔH_{k-1} is defined by the following conditions:

$$\begin{aligned} \Delta H_{k-1} e_j &= 0, \quad 0 \leq j \leq k - 2, \\ \Delta H_{k-1} e_{k-1} &= a r_{k-1} - H_{k-1} e_{k-1}. \end{aligned} \quad (4.27)$$

The latter equality will evidently be satisfied if we assume

$$\Delta H_{k-1} = a \frac{r_{k-1} u_{k-1}^*}{(u_{k-1}, e_{k-1})} - \frac{H_{k-1} e_{k-1} v_{k-1}^*}{(v_{k-1}, e_{k-1})} \quad (4.28)$$

where u_{k-1}, v_{k-1} are unknown vectors. It is necessary that the vectors be such that the first of the conditions (4.27) is satisfied, i.e.

$$(u_{k-1}, e_j) = 0, \quad (v_{k-1}, e_j) = 0, \quad 0 \leq j \leq k - 2. \quad (4.29)$$

Clearly, vectors u_{k-1}, v_{k-1} must also satisfy conditions

$$(u_{k-1}, e_{k-1}) \neq 0, \quad (v_{k-1}, e_{k-1}) \neq 0. \quad (4.30)$$

Taking into account (4.23) it is clear that conditions (4.29) will be satisfied if we choose $u_{k-1} = v_{k-1} = r_{k-1}$. Conditions (4.30) will also be satisfied since

$$(r_{k-1}, e_{k-1}) = (r_{k-1}, A r_{k-1}) > 0 \quad (4.31)$$

according to the properties of matrix A .

Vectors u_{k-1} , v_{k-1} can also be chosen by using the following considerations. If condition (4.20) is satisfied, then we have

$$(Ap_{k-1}, p_j) = \left(\frac{1}{\alpha_{k-1}} e_{k-1}, \frac{1}{\alpha_j} r_j \right) = 0, \quad 0 \leq j \leq k-2.$$

Making use of (4.25) we have then

$$a(e_{k-1}, r_j) = (e_{k-1}, H_{k-1}e_j) = (H_{k-1}^* e_{k-1}, e_j) = 0, \\ 0 \leq j \leq k-2.$$

It follows that in order to satisfy (4.29) we can assume

$$u_{k-1} = v_{k-1} = H_{k-1}^* e_{k-1}.$$

In general, if we choose vectors u_{k-1} and v_{k-1} in the form

$$u_{k-1} = t_{1,k} r_{k-1} + t_{2,k} H_{k-1}^* e_{k-1}, \\ v_{k-1} = t_{3,k} r_{k-1} + t_{4,k} H_{k-1}^* e_{k-1} \quad (4.32)$$

where $t_{1,k}$, $t_{2,k}$, $t_{3,k}$, $t_{4,k}$ are arbitrary numbers (which in principle can change with changing k), then conditions (4.29) will obviously be satisfied. In order to satisfy conditions (4.30) the quantities $t_{i,k}$, $i = 1, \dots, 4$ should be adjusted, if necessary (in particular, as has been noted, conditions (4.30) will be satisfied with $t_{1,k} = t_{3,k} = 1$, $t_{2,k} = t_{4,k} = 0$; see (4.31)).

Thus, choosing vectors u_{k-1} , v_{k-1} in the form (4.32) we are able to construct matrix ΔH_{k-1} by formula (4.28) and in this way establish the recursive relations for constructing such a matrix H_k that the vector p_k which it determines will satisfy the conditions of A -orthogonality (4.20). To each pair of vectors u_{k-1} , v_{k-1} and constant a chosen there will correspond their particular matrix ΔH_{k-1} and, consequently, matrix H_k . In other words, with different vectors u_k , v_k and constant a we shall obtain different algorithms for constructing conjugate vectors, i.e. shall construct different methods of conjugate directions.

General Properties of the Methods

Let us try to establish the general properties of methods of conjugate directions, which can be constructed in the manner described above.

First of all, it is necessary to ascertain whether condition (4.18) is satisfied by the methods under consideration since in working out the methods of constructing the algorithms, it is assumed that this condition is fulfilled.

Another interesting question is whether the directions p_j , $j = 0, 1, \dots, n-1$ which are determined by different matrices H_j

differ from one another, i.e. whether points $x_1, \dots, x_{n-1}$ are different for different algorithms (on condition that point x_0 remains the same) or they coincide.

In order to answer these questions let us present vector $-p_j = -H_j^* f_j$ using the recursive formula for matrix H_j and expressions (4.28), (4.32) in the following form:

$$H_j^* f_j = (H_{j-1} + \Delta H_{j-1})^* f_j.$$

Making use of (4.24) we can write

$$\begin{aligned} \Delta H_{j-1}^* f_j &= - \frac{v_{j-1} (e_{j-1}, H_{j-1}^* f_j)}{(v_{j-1}, e_{j-1})} \\ &= - \frac{(t_{3,j} r_{j-1} + t_{4,j} H_{j-1}^* e_{j-1}) (e_{j-1}, H_{j-1}^* f_j)}{(v_{j-1}, e_{j-1})}. \end{aligned}$$

If we also take into account, that

$$H_{j-1}^* e_{j-1} = H_{j-1}^* f_j - H_{j-1}^* f_{j-1} = H_{j-1}^* f_j + p_{j-1},$$

then vector $-p_j$ can be written in the following form:

$$\begin{aligned} H_j^* f_j &= H_{j-1}^* f_j \left[1 - \frac{t_{4,j} (e_{j-1}, H_{j-1}^* f_j)}{(v_{j-1}, e_{j-1})} \right] \\ &\quad - \frac{t_{3,j} r_{j-1} e_{j-1}^* + t_{4,j} p_{j-1} e_{j-1}^*}{(v_{j-1}, e_{j-1})} H_{j-1}^* f_j \\ &= \left[I \left(1 - \frac{t_{4,j} (e_{j-1}, H_{j-1}^* f_j)}{(v_{j-1}, e_{j-1})} \right) - \frac{r_{j-1} e_{j-1}^*}{(v_{j-1}, e_{j-1})} \left(t_{3,j} + \frac{t_{4,j}}{\alpha_{j-1}} \right) \right] H_{j-1}^* f_j. \end{aligned}$$

Further

$$\begin{aligned} (v_{j-1}, e_{j-1}) &= (t_{3,j} r_{j-1} + t_{4,j} H_{j-1}^* e_{j-1}, e_{j-1}) \\ &= t_{3,j} (r_{j-1}, e_{j-1}) + t_{4,j} (e_{j-1}, H_{j-1}^* f_j) + t_{4,j} (e_{j-1}, p_{j-1}) \end{aligned} \quad (4.33)$$

hence

$$\frac{(r_{j-1}, e_{j-1})}{(v_{j-1}, e_{j-1})} \left(t_{3,j} + \frac{t_{4,j}}{\alpha_{j-1}} \right) = 1 - \frac{t_{4,j} (e_{j-1}, H_{j-1}^* f_j)}{(v_{j-1}, e_{j-1})}.$$

Using this expression to transform the formula for $H_j^* f_j$, we obtain

$$H_j^* f_j = \gamma_j \left(I - \frac{r_{j-1} e_{j-1}^*}{(r_{j-1}, e_{j-1})} \right) H_{j-1}^* f_j \quad (4.34)$$

where

$$\gamma_j = 1 - \frac{t_{4,j} (e_{j-1}, H_{j-1}^* f_j)}{(v_{j-1}, e_{j-1})}.$$

If vector v_{j-1} satisfies condition (4.30) and $\alpha_{j-1} t_{3,j} \neq -t_{4,j}$, with any $j = 1, 2, \dots$ factor $\gamma_j \neq 0$ since

$$\frac{t_{4,j} (e_{j-1}, H_{j-1}^* f_j)}{(v_{j-1}, e_{j-1})} \neq 1.$$

This inequality is easily checked by comparing the numerator of the ratio with expression (4.33). Further, suppose that factors $t_{3,j}$ and $t_{4,j}$ are such that with $j \geq 1$ the conditions $(v_{j-1}, e_{j-1}) \neq 0$ and $\gamma_j \neq 0$ be satisfied.

Let us perform the scalar multiplication of the two sides of equality (4.34) by f'_k

$$(f'_k, H_j^* f'_j) = \gamma_j \left[(f'_k)^* - \frac{(f'_k, r_{j-1}) e_{j-1}^*}{(r_{j-1}, e_{j-1})} \right] H_{j-1}^* f'_j. \quad (4.35)$$

Since $\gamma_j \neq 0$ and $(f'_k, H_j^* f'_j) = 0$, $(f'_k, r_j) = 0$ with $j \leq k-1$ (by (4.21) and (4.24)), it follows from (4.35) that

$$(f'_k, H_{j-1}^* f'_j) = 0, \quad 1 \leq j \leq k-1. \quad (4.36)$$

Subtracting the equalities $(f'_k, H_{j-1}^* f'_{j-1}) = 0$, $1 \leq j \leq k-1$ from (4.36) we obtain:

$$(f'_k, H_i^* e_i) = 0, \quad 0 \leq i \leq k-2. \quad (4.37)$$

Let us now prove using the relations obtained that with $0 \leq i \leq j-2$ equalities $H_{i+1}^* f'_j = H_i^* f'_j$ hold, i.e.

$$H_{j-1}^* f'_j = H_{j-2}^* f'_j = \dots = H_0^* f'_j. \quad (4.38)$$

Using the recursive formula for matrix H_i and taking into account conditions (4.24), we can write vector $H_{i+1}^* f'_j$ in the following form:

$$H_{i+1}^* f'_j = H_i^* f'_j - \frac{v_i (H_i e_i, f'_j)}{(v_i, e_i)}, \quad 0 \leq i \leq j-2. \quad (4.39)$$

In order to prove equalities (4.38) it is necessary to show that

$$(f'_j, H_i e_i) = 0, \quad 0 \leq i \leq j-2. \quad (4.40)$$

Using again the recursive formula for H_i we obtain:

$$(f'_j, H_{i+1} e_{i+1}) = (f'_j, H_0 e_{i+1}) - \sum_{s=0}^i \frac{(f'_j, H_s e_s) (v_s, e_{i+1})}{(v_s, e_s)}, \quad 0 \leq i \leq j-2, \quad (4.41)$$

$$(f'_j, H_{i+1}^* e_{i+1}) = (f'_j, H_0^* e_{i+1}) - \sum_{s=0}^i \frac{(v_s, f'_j) (H_s e_s, e_{i+1})}{(v_s, e_s)}, \quad 0 \leq i \leq j-2. \quad (4.42)$$

Because of conditions (4.24) and (4.37), we have

$$(v_i, f'_j) = t_{3,i} (r_i, f'_j) + t_{4,i} (H_i^* e_i, f'_j) = 0, \quad 0 \leq i \leq j-2. \quad (4.43)$$

Taking into account equalities (4.43) and conditions (4.37), it follows from (4.42) that

$$(f'_j, H_0 e_{i+1}) = 0, \quad 0 \leq i \leq j-3. \quad (4.44)$$

Let us now consider the relations (4.41). With $i = 0$ we have by (4.37) $(H_0 e_0, f'_j) = (H_0^* e_0, f'_j) = 0$ and by (4.44) $(f'_j, H_0 e_1) = 0$. Consequently, $(H_1 e_1, f'_j) = 0$. Further, we similarly establish that equalities (4.38) hold.

Taking into account these equalities, we can write expression (4.34) in the following form:

$$H_j^* f'_j = \gamma_j \left(I - \frac{r_{j-1} e_{j-1}^*}{(r_{j-1}, e_{j-1})} \right) H_0 f'_j. \quad (4.45)$$

This is the formula that enables us to answer the questions formulated at the beginning of this subsection.

By scalar multiplication of the two sides of (4.45) by f'_j and using condition (4.24), we obtain

$$-(f'_j, p_j) = \gamma_j (f'_j, H_0 f'_j), \quad j \geq 0. \quad (4.46)$$

If H_0 is a strictly positive definite matrix, then $(f'_j, H_0 f'_j) > 0$. Consequently, if $\gamma_j \neq 0$, then it follows from (4.46) that $(f'_j, p_j) \neq 0$.

Thus, it follows from (4.46) that the assumption that condition (4.18) can be satisfied, used in working out the methods of constructing conjugate vectors, proves to hold if H_0 is a symmetric, strictly positive definite matrix.

In order to ascertain whether vectors p_i and points x_{i+1} , $i = 0, 1, \dots, n-1$ are different in different algorithms we turn again to formula (4.45).

The first step in any method of conjugate directions is the same (given the same matrix H_0) since $x_1 = x_0 - \alpha_0 H_0^* f'_0$ and α_0 is chosen under the condition $\min_{\alpha} f(x_0 - \alpha H_0^* f'_0)$. Consequently, point x_1 and therefore vectors r_0, e_0, f'_1 will also be the same in any algorithm which can be constructed by the method described. But then as it follows from (4.45), the direction

$$p_1 = -H_1^* f'_1 = -\gamma_1 \left[I - \frac{r_0 e_0^*}{(r_0, e_0)} \right] H_0 f'_1$$

also will not depend on the choice of vectors u_0, v_0 (which satisfy the requirements formulated), i.e. will not depend on the method of constructing matrix H_1 . To be more precise, vectors p_1 obtained with different methods of constructing matrix H_1 will differ only by the scalar factor γ_1 . However since the quantity α_1 is chosen under the condition $\min_{\alpha} f(x_1 + \alpha p_1)$, point x_2 which provides the minimum will be the same whatever the method of constructing matrix

H_1 (by virtue of the strict convexity of $f(x)$). Consequently, the quantities r_1, e_1, f'_2 will be the same for different methods of conjugate directions. Continuing this argument based on expressing vector p_k by formula (4.45) we conclude that points $x_0, x_1, \dots, x_n$ are independent of the choice of vectors u_k, v_k , i.e. of the method of constructing matrix H_k . Thus the successive approximations to the solution of the problem of minimization of a quadratic function are the same for different methods of conjugate directions.

One more remark.

It could be noted above that the first of the two matrices that form matrix ΔH_j (4.28), $j = 0, 1, \dots, k-1$ takes no part in constructing vector p_k . Indeed in determining vector $\Delta H_j^* f'_k$ we find according to conditions (4.24) that

$$a \frac{u_j(r_j, f'_k)}{(u_j, e_j)} = 0, \quad 0 \leq j \leq k-1,$$

i.e. matrices

$$a \frac{r_j u_j^*}{(u_j, e_j)}, \quad 0 \leq j \leq k-1$$

take no part in constructing vector $-p_k = H_k^* f'_k =$
 $= \left(H_0 + \sum_{j=0}^{k-1} \Delta H_j \right)^* f'_k$. However, they affect considerably the properties of matrix H_k , in particular the properties of matrix H_n . We shall take notice of this fact in studying concrete algorithms in the following subsection. Here we note only that the difference in the properties of matrix H_n tells on the properties of the methods of conjugate directions in the minimization of nonquadratic functions.

Concrete Algorithms

Let us now consider several formulas which can be used in constructing conjugate directions. Let us repeat that each of such formulas determines a method of conjugate directions consisting in constructing successive approximations to the solution by formulas $x_{k+1} = x_k + \alpha_k p_k$, $p_k = -H_k^* f'_k$, $k = 0, 1, \dots, n-1$ (4.47) where α_k is chosen under the condition $\min_{\alpha} f(x_k + \alpha p_k)$, i.e. is determined by expressions (4.12) or (4.16).

(1) We set in (4.28) $a = 1$, $u_{k-1} = r_{k-1}$, $v_{k-1} = H_{k-1}^* e_{k-1}$ (i.e. in formulas (4.32) $t_{1,k} = t_1 = 1$, $t_{2,k} = t_2 = 0$, $t_{3,k} = t_3 = 0$, $t_{4,k} = t_4 = 1$). Then

$$H_k = H_{k-1} + \frac{r_{k-1} r_{k-1}^*}{(r_{k-1}, e_{k-1})} - \frac{H_{k-1} e_{k-1} e_{k-1}^* H_{k-1}}{(H_{k-1}^* e_{k-1}, e_{k-1})}. \quad (4.48)$$

Let us study some properties of matrix H_k obtained by this method.

Matrix H_k is symmetric. This fact is easily established by induction. Matrix H_0 is symmetric. The two matrices which form ΔH_0 are symmetric too (the second one by virtue of the symmetry of H_0). Therefore, H_1 is a symmetric matrix. Similar arguments hold for any $k = 2, \dots, n$.

Matrix H_k is strictly positive definite. We give a proof by induction. Matrix H_0 is strictly positive definite. Let H_k be a strictly positive definite matrix. Then for any $x \in E^n$

$$\begin{aligned} (H_{k+1}x, x) &= (H_k x, x) + \frac{(r_k, x)^2}{(r_k, e_k)^2} - \frac{(H_k e_k, x)^2}{(H_k e_k, e_k)} \\ &= \frac{(H_k x, x)(H_k e_k, e_k) - (H_k e_k, x)^2}{(H_k e_k, e_k)} + \frac{(r_k, x)^2}{(r_k, e_k)}. \end{aligned}$$

By hypothesis, there is a square root $H_k^{1/2}$ (D. K. Faddeev and V. N. Faddeeva). Consequently taking into account the symmetry of matrix H_k , we have

$$(H_k x, x) = (H_k^{1/2} H_k^{1/2} x, x) = (H_k^{1/2} x, H_k^{1/2} x) = (y, y);$$

similarly

$$(H_k e_k, e_k) = (H_k^{1/2} e_k, H_k^{1/2} e_k) = (z, z),$$

$$(H_k e_k, x) = (H_k^{1/2} e_k, H_k^{1/2} x) = (z, y).$$

Making use of these relations and applying Cauchy-Buniakowski's inequality we conclude that the following inequality holds:

$$(H_k x, x)(H_k e_k, e_k) - (H_k e_k, x)^2 = (y, y)(z, z) - (z, y)^2 \geq 0$$

and this inequality holds only if $z = y$, i.e. since H_k is nonsingular, only if $x = e_k$. But in this case $(r_k, x) = (r_k, e_k) = (r_k, A r_k) > 0$. Thus for any $x \neq 0$, we have

$$(H_{k+1}x, x) = \frac{(y, y)(z, z) - (z, y)^2}{(H_k e_k, e_k)} + \frac{(r_k, x)^2}{(r_k, e_k)} > 0$$

and this proves that our reasoning by induction holds.

Matrix $H_n = A^{-1}$. Indeed, H_k satisfies (4.25) with $a = 1$, i.e.

$H_n e_j = r_j$, $j = 0, 1, \dots, n-1$, or making use of (4.19)

$$H_n A r_j = r_j, \quad j = 0, 1, \dots, n-1.$$

It follows that vectors $r_0, \dots, r_{n-1}$ are eigenvectors of matrix $H_n A$ with eigenvalues equal to unity. Hence, taking into account the linear independence of conjugate vectors $r_i, i = 0, 1, \dots, n-1$ we have $H_n A = I$, i.e.

$$H_n = A^{-1}.$$

But from (4.8)

$$A^{-1} = \sum_{i=0}^{n-1} \frac{r_i r_i^*}{(A r_i, r_i)} = \sum_i \frac{r_i r_i^*}{(e_i, r_i)},$$

i.e. we find that matrix H_n is determined only by the matrices

$$\frac{r_i u_i^*}{(r_i, e_i)} = \frac{r_i r_i^*}{(r_i, e_i)}$$

(this was mentioned at the end of the preceding subsection).

(2) Another method of constructing H_k is obtained if we take $a = 1$ in (4.28) and choose $u_{k-1} = v_{k-1} = r_{k-1}$. We have then

$$H_k = H_{k-1} + (r_{k-1} - H_{k-1} e_{k-1}) \frac{r_{k-1}^*}{(r_{k-1}, e_{k-1})}. \quad (4.49)$$

Matrix H_k thus determined is no more symmetric. Since $a = 1$, we have now $H_n = A^{-1}$ and this can be demonstrated just in the same way as for method (4.48).

Using (4.49) we can obtain a somewhat different formula for constructing H_k . We write (4.49) in the following form:

$$H_k = H_0 + \sum_{i=0}^{k-1} (r_i - H_i e_i) \frac{r_i^*}{(r_i, e_i)}. \quad (4.50)$$

According to the conditions of conjugateness (4.20) (taking into account formulas (4.19)), we have $(e_k, r_j) = 0, 0 \leq j \leq k-1$. Consequently, it follows from (4.50) that

$$H_k e_k = H_0 e_k, \quad k = 0, 1, \dots, n-1. \quad (4.51)$$

Thus formula (4.49) can be written as follows:

$$H_k = H_{k-1} + (r_{k-1} - H_0 e_{k-1}) \frac{r_{k-1}^*}{(r_{k-1}, e_{k-1})}. \quad (4.52)$$

If $H_0 = I$, this formula proves somewhat simpler than (4.49).

(3) Let us choose $a = 0, v_{k-1} = r_{k-1}$. In this case

$$H_k = H_{k-1} - \frac{H_{k-1} e_{k-1} r_{k-1}^*}{(r_{k-1}, e_{k-1})}. \quad (4.53)$$

With $a = 0$ it follows from (4.25) that $H_n e_j = 0, j = 0, 1, \dots, n-1$. Since vectors $e_0, \dots, e_{n-1}$ are linearly independent,

these equalities imply that $H_n = 0$ (the linear independence of vectors $e_i = Ar_i$, $i = 0, 1, \dots, n-1$ follows from the linear independence of conjugate vectors r_i and the properties of matrix A).

Since condition (4.51) is satisfied for formula (4.53) too, the latter can be written in the following form:

$$H_k = H_{k-1} - \frac{H_0 e_{k-1} r_{k-1}^*}{(r_{k-1}, e_{k-1})}. \quad (4.54)$$

The constructing of methods of conjugate directions can be continued by choosing various combinations of constant a and vectors u_k, v_k by formulas (4.32) but we shall not do so (there and further, speaking of a concrete method, for example (4.48), we have in mind method (4.47) in which formula (4.48) is used for constructing matrix H_k).

Let us make a remark here. Strictly speaking, it was necessary in each of the methods treated above to check whether conditions (4.30) were satisfied by vectors u_k, v_k . It is easy to ascertain that in all the methods studied, these conditions were satisfied. For instance, in the case $u_k = v_k = r_k$ the satisfying of conditions (4.30) was already mentioned in the subsection on p. 85. In method (4.48) $v_k = H_k^* e_k$; however, since matrix H_k is positive definite, we have $(v_k, e_k) = (H_k e_k, e_k) > 0$, i.e. conditions (4.30) are also satisfied. Thus, in accordance with the results of the subsection on p. 89, condition (4.18) is satisfied by the methods discussed, i.e. the methods are guaranteed to be nondegenerate.

Let us now derive formulas directly applicable to the calculation of vectors p_k defined by different matrices H_k . This is easily done by using formula (4.45). Since $r_{k-1} = \alpha_{k-1} p_{k-1}$, we have from (4.45)

$$p_k = -\gamma_k (H_0 f'_k - \beta_k p_{k-1}) \quad (4.55)$$

where

$$\beta_k = \frac{(H_0 f'_k, e_{k-1})}{(p_{k-1}, e_{k-1})}. \quad (4.56)$$

If $v_k = H_k^* e_k$, then

$$\gamma_k = 1 - \frac{(e_{k-1}, H_{k-1}^* f'_k)}{(H_{k-1} e_{k-1}, e_{k-1})}.$$

According to equalities (4.38) we have $(e_{k-1}, H_{k-1}^* f'_k) = (e_{k-1}, H_0 f'_k)$. Further, because of (4.21) and (4.38), $(H_k e_k, e_k) = (H_k f'_{k+1}, f'_{k+1}) + (H_k, f'_k) = (H_0 f'_{k+1}, f'_{k+1}) - (p_k, f'_k)$. The equalities obtained show that

$$\gamma_k = 1 - \frac{(H_0 f'_k, e_{k-1})}{(H_0 f'_k, f'_k) - (p_{k-1}, f'_{k-1})}. \quad (4.57)$$

Note that from (4.45), because of (4.21) and (4.24), it follows that (provided $\gamma_j \neq 0$)

$$(f'_k, H_0 f'_j) = 0, \quad 0 \leq j \leq k-1. \quad (4.58)$$

Taking this into account, it proves that

$$(H_0 f'_k, e_{k-1}) = (H_0 f'_k, f'_k). \quad (4.59)$$

Using equalities (4.59) and (4.57) we find that

$$\gamma_k = - \frac{(p_{k-1}, f'_{k-1})}{(H_0 f'_k, f'_k) - (p_{k-1}, f'_{k-1})}. \quad (4.60)$$

Note also that

$$(f'_k, p_k) = (f'_k, p_k) - (f'_{k+1}, p_k) = - (e_k, p_k). \quad (4.61)$$

Comparing formulas (4.56) and (4.60) and taking into account (4.59) and (4.61), it is easy to establish that $\gamma_k = \frac{1}{1+\beta_k}$. Hence $\gamma_k \beta_k = 1 - \gamma_k$. Consequently, formula (4.55) which determines vector p_k , in the case where in constructing matrix H_k we use vector $v_{k-1} = H_{k-1}^* e_{k-1}$, can be written in the form

$$p_k = -\gamma_k H_0 f'_k + (1 - \gamma_k) p_{k-1} \quad (4.62)$$

where coefficient γ_k is determined by one of the formulas (4.57) or (4.60). Vector p_k can be written in the alternative form:

$$p_k = -H_0 f'_k + \tilde{\beta}_k (H_0 f'_k + p_{k-1}) \quad (4.63)$$

where

$$\tilde{\beta}_k = 1 - \gamma_k = \frac{(H_0 f'_k, f'_k)}{(H_0 f'_k, f'_k) - (p_{k-1}, f'_{k-1})}. \quad (4.64)$$

Other expressions can be obtained for coefficient $\tilde{\beta}_k$ if equalities (4.59), (4.61), (4.46) are made use of; note that the last of these formulas can be written in the form

$$(f'_k, p_k) = (\tilde{\beta}_k - 1) (H_0 f'_k, f'_k). \quad (4.65)$$

Calculating in expressions (4.62), (4.63) coefficients γ_k , $\tilde{\beta}_k$ by different formulas, we practically obtain different methods of conjugate directions. It should be stressed that in the minimization of non-quadratic functions different formulas for p_k determine different vectors (both as to magnitude and as to direction). Especially simple is the construction of vector p_k using expression (4.45) if vector $v_{k-1} = r_{k-1}$ is chosen in constructing matrix H_k . In this case $t_{k,k} = 0$, therefore $\gamma_k = 1$, and from (4.55) we obtain

$$p_k = -H_0 f'_k + \beta_k p_{k-1} \quad (4.66)$$

where β_k is calculated by formula (4.56). If we use equalities (4.59), (4.61) and (4.46) (the last one in the case under consideration takes the form

$$(p_k, f'_k) = - (H_0 f'_k, f'_k) \quad (4.67)$$

then for determining coefficient β_k one of the following formulas can be obtained:

$$\beta_k = - \frac{(H_0 f'_k, e_{k-1})}{(p_{k-1}, f'_{k-1})} = - \frac{(H_0 f'_k, f'_k)}{(p_{k-1}, f'_{k-1})} = \frac{(H_0 f'_k, f'_k)}{(H_0 f'_{k-1}, f'_{k-1})} \quad (4.68)$$

Expressions (4.62), (4.63), (4.66) which determine vector p_k in their turn can be given the form $p_k = -H_k^* f'_k$, where matrix H_k depends on the coefficients γ_k , $\tilde{\beta}_k$, β_k . For instance, if coefficient $\tilde{\beta}_k$ in (4.63) is calculated by formula (4.64), then the corresponding matrix

$$H_k = H_0 - \frac{H_0 f'_k (H_0 f'_k + p_{k-1})^*}{(H_0 f'_k, f'_k) - (p_{k-1}, f'_{k-1})} \quad (4.69)$$

where $H_n = H_0$ (since $f'_n = 0$). If in (4.66) β_k is calculated by the first of formulas (4.68), then vector p_k is determined by matrix

$$H_k = H_0 + \frac{H_0 e_{k-1} p_{k-1}^*}{(p_{k-1}, f'_{k-1})} \quad (4.70)$$

and if use is made of the second of formulas (4.68), then

$$H_k = H_0 + \frac{H_0 f'_k p_{k-1}^*}{(p_{k-1}, f'_{k-1})}. \quad (4.71)$$

Note that in formula (4.71) $H_n = H_0$ (since $f'_n = 0$); in (4.70) $H_n \neq H_0$.

The reader himself can obtain other formulas for constructing H_k .

The simplest formula for calculating A -orthogonal vectors can be obtained by choosing $H_0 = I$ in (4.66). In this case

$$p_k = -f'_k + \beta_k p_{k-1} \quad (4.72)$$

where β_k is determined, for instance, by one of the following formulas:

$$\beta_k = - \frac{(f'_k, e_{k-1})}{(p_{k-1}, f'_{k-1})} = - \frac{(f'_k, f'_k)}{(p_{k-1}, f'_{k-1})} = \frac{(f'_k, f'_k)}{(f'_{k-1}, f'_{k-1})}. \quad (4.73)$$

Method (4.47) in which conjugate vectors are constructed by (4.72), (4.73) is widely known as the method of conjugate gradients (this name is due to conditions (4.58)).

Minimization of a Convex Quadratic Function

Until now we considered methods of A -orthogonal directions for the minimization of a strictly convex quadratic function, i.e. assumed matrix A to be strictly positive definite.

Let now the function

$$f(x) = \frac{1}{2} (Ax, x) + (b, x) + c$$

be convex, i.e. matrix A be positive definite: $(Ax, x) \geq 0$ with any $x \neq 0$. Suppose that this function has a minimum.

Let us study the problem of the application of methods of conjugate directions in this case. Consider preliminarily certain properties of function $f(x)$.

(1) *If $(Ap, p) = 0$ then of necessity*

$$Ap = 0. \quad (4.74)$$

In fact if $(Ap, p) = 0$, then p is the minimum point of the convex function $\varphi(x) = \frac{1}{2} (Ax, x)$. But at the minimum point the necessary condition of an extremum

$$\varphi'(p) = Ap = 0$$

must be satisfied.

(2) *If p is the minimum point of the convex function $\varphi(x) = \frac{1}{2} (Ax, x)$, then of necessity*

$$(b, p) = 0. \quad (4.75)$$

In fact, if $(Ap, p) = 0$ and $(b, p) > 0$, then $f(\alpha p) = \alpha(b, p) + c \rightarrow -\infty$ as $\alpha \rightarrow -\infty$, i.e. $f(x)$ does not attain the minimum and this contradicts the assumption.

The case where $(b, p) < 0$ is treated in a similar manner.

(3) *The minimum point of function $f(x)$ is not unique.*

Indeed, any minimum point of a convex quadratic function $f(x)$ must be a solution of the linear system $Ax + b = 0$ and conversely, since the condition $f'(x) = Ax + b = 0$ is a necessary and (since there is a minimum of $f(x)$) sufficient condition for an extremum of the convex function $f(x)$ (corollary 3.2 of Chap. I). However, the rank of matrix A is lower than the number of unknowns (condition $(Ax, x) \geq 0$ means that matrix A is singular, see (4.74)) and so the system $Ax + b = 0$ has no unique solution.

(4) *If $(Ap, p) = 0$ and $z \in E^n$ is an arbitrary point, then of necessity*

$$(f'(z), p) = 0. \quad (4.76)$$

Indeed, according to conditions (4.74) and (4.75)

$$(f'(z), p) = (Az + b, p) = (Ap, z) + (b, p) = 0.$$

Equality (4.76) can be interpreted as follows. The set of solutions of the problem of the minimization of the function $\varphi(x)$ forms an $n - q$ -dimensional hyperplane, where q is the rank of matrix A . (This hyperplane belongs to the level surface of function $f(x)$ since if p is an arbitrary point of minimum of $\varphi(x) = \frac{1}{2}(Ax, x)$ then using (4.75) we have

$$f(p) = \frac{1}{2}(Ap, p) + (b, p) + c = c.)$$

Consequently, equality (4.76) means that the gradient of function $f(x)$ at any point is situated in a q -dimensional subspace which is orthogonal to plane $Ap = 0$. It follows that the number of linearly independent vectors $f'(x)$ is equal to $q < n$ (for a convex quadratic function $f(x)$).

Having in mind the above properties of function $f(x)$ let us again turn to the problem of the application of methods of conjugate directions to the solving of the problem under consideration.

For simplicity, suppose that $H_0 = I$ and consider method (4.72). We denote the subspace which contains vectors $f'(x)$ by E^q . It is easy to ascertain that vector p_k defined by formula (4.72) belongs to E^q . Indeed, $p_0 = -f'_0 \in E^q$ and consequently with any k , vector p_k is a linear combination of vectors which belong to subspace E^q . Consequently, the function minimization (process (4.47)) by method (4.72) is practically performed in subspace E^q . Now in this subspace, condition $(Ax, x) > 0$ is satisfied for any $x \neq 0$. This means that owing to space E^q being finite-dimensional we have for any $x \in E^q$

$$m_1 \|x\|^2 \leq (Ax, x) \leq M_1 \|x\|^2, \quad m_1 > 0, \quad M_1 \leq M.$$

Hence function $f(x)$ is strictly convex in the subspace under consideration, therefore all of the properties of methods of conjugate directions discussed in the preceding subsections hold also in our case. In particular, equalities (4.58) which show that vectors f'_i , $i = 0, 1, \dots, k$ are linearly independent hold in our case. However, there cannot be more than q linearly independent vectors in subspace E^q . Consequently, with a certain $k \leq q - 1$ the process of constructing conjugate vectors must be truncated. Since the method is nondegenerate (property (4.18)), this will occur only if $f'_k = 0$.

It is clear from what has been stated above that in minimizing a function by method (4.72) at a certain $k \leq q - 1$ we shall of necessity find that $f'_k = 0$. Since directions p_k determined by different methods of conjugate directions coincide (to an accuracy of a scalar factor), everything stated above holds not only for method

(4.72) but also for other algorithms studied in this section. By a somewhat more complicated reasoning it can be shown that the result obtained holds also in the case where H_0 is an arbitrary strictly positive definite matrix.

Thus, *methods of conjugate directions make it possible to find the minimum point of a convex quadratic function and the solution of the problem is obtained after less than n steps.*

We suppose now that the convex function $f(x)$ does not attain the minimum; this will be the case, as can be seen from the proof of equality (4.75), if an arbitrary vector p which minimizes function $\varphi(x) = (Ax, x)$ is such that $(b, p) \neq 0$.

Let us consider what the processes of constructing conjugate directions lead to in this case.

We first describe one property of A -orthogonal vectors which takes place under the condition that $f(x)$ is a convex function (but not strictly convex).

If matrix A is positive definite, then at least one of the conjugate vectors p_k , $0 \leq k \leq n - 1$ satisfies the condition

$$(Ap_k, p_k) = 0. \quad (4.77)$$

That this statement holds is shown as follows.

If the conditions

$$(Ap_i, p_i) > 0 \quad (4.78)$$

were satisfied for any $i = 0, 1, \dots, n - 1$, then the system of vectors $p_0, p_1, \dots, p_{n-1}$ would be linearly independent. Indeed, suppose that conditions (4.78) are satisfied and that

$$\sum_{i=0}^{n-1} \delta_i p_i = 0$$

where, for example, $\delta_0 \neq 0$. Then by (scalar) multiplication of both sides of the equality by Ap_0 we obtain on the left-hand side $\delta_0 (Ap_0, p_0) > 0$, i.e. the result is a contradiction. Consequently, with conditions (4.78), satisfied vectors $\{p_i\}_0^{n-1}$ would form a basis in E^n and therefore any vector z could be written in the form

$$z = \sum_{i=0}^{n-1} a_i p_i$$

where at least one of the coefficients $a_i \neq 0$. However, in this case we would have

$$(Az, z) = (A \sum_i a_i p_i, \sum_i a_i p_i) = \sum_i a_i^2 (Ap_i, p_i) > 0.$$

This contradicts condition $(Ax, x) \geq 0$.

Thus the initial assumption that conditions (4.78) are fulfilled is false.

According to the property of A -orthogonal vectors discussed, in applying any method of conjugate directions with a certain $k \geq 0$ we shall find equality (4.77) satisfied. Since for the function under consideration properties (4.75), (4.76) do not take place, parameter α_k , with (4.77) satisfied, which is calculated by formula (4.16) will become infinite, i.e. the further constructing of conjugate directions will prove impossible.

Discussion of Results

Thus, we have considered a general scheme of constructing methods of conjugate directions and on its basis obtained many concrete algorithms. Any of these methods makes it possible to find the minimum of a convex quadratic function after a number of steps in process (4.47) not exceeding n . Besides, we have made it clear that successive approximations to the solution $x_0, x_1, \dots, x_k$ obtained by using different algorithms prove to be identical.

If algorithms are judged by the amount of calculations per iteration, then algorithms (4.62), (4.63), (4.66) should certainly be preferred. These methods are especially easy to implement if the identity matrix I is chosen as the initial matrix H_0 and it seems that for most problems the choice $H_0 = I$ is the most expedient.

In this case methods (4.63), (4.66) by the work per iteration differ but slightly from the gradient method but are considerably more effective owing to the fact that process (4.47) proves to have a finite number of steps.

The advantage of methods (4.62), (4.63) and (4.66) when implemented on computers is that they require but a slightly larger storage capacity of the computer than that required by the methods of steepest descent.

Methods of conjugate directions in which in order to determine the direction of motion matrices ((4.48), (4.49), (4.52)-(4.54)) are preliminarily constructed are somewhat inferior to methods (4.63), (4.66) in the aspects considered; however, they retain their considerable advantage over gradient methods. All the methods of the class under consideration have advantage over Newton's method in that they do not require the calculation of second derivatives of the function.

The question can be raised whether it is worth while to consider methods where we use preliminary construction of matrices if they are inferior to methods (4.63), (4.72) in the computational effort and computer memory required.

However, it should be kept in mind that we are making a purely theoretical estimate of the methods and do not take into account such an important factor as the sensitiveness of an algorithm to errors in computations. This factor can change considerably the relation between the amounts of computations involved in solving the problem by different algorithms. It should also be noted that methods (4.48), (4.49), (4.52), for instance, in solving a minimization problem allow to obtain at the same time the inverse matrix A^{-1} and this may be useful in some cases.

The difference in properties of algorithms tells considerably when they are used for minimization of nonquadratic functions; this will be discussed in the next section.

Methods of conjugate directions prove useful in one more aspect; they make it possible to establish whether the sign of the matrix is fixed. Thus according to the results of the subsection on p. 99, if matrix A is positive definite and function $f(x)$ does not attain the minimum, then at a certain step we shall have $\alpha_k = \infty$. However, if matrix A is not positive definite, we find at a certain step of process (4.47) that $\alpha_k < 0$. Thus the value of parameter α_k determines the sign of matrix A .

The effectiveness of methods of conjugate directions is the reason of their more and more extensive application to the minimization of quadratic functions and solution of systems of linear equations.

5. METHODS OF CONJUGATE DIRECTIONS. MINIMIZATION OF ARBITRARY FUNCTIONS

Considerations about the Applicability of the Methods

Suppose that we intend to make use of the process

$$x_{k+1} = x_k + \alpha_k p_k, \quad p_k = -H_k^* f'_k, \quad k = 0, 1, \dots, \quad (5.1)$$

where vector p_k (or matrix H_k) is determined by one of the methods studied in the preceding section, for the minimization of an arbitrary (not quadratic) convex function $f(x)$. In this case, matrix $f''(x)$ will have different elements at different points of sequence (5.1); by virtue of this fact vectors $p_0, \dots, p_k$ constructed by any of the methods described in the subsection on p. 93 will not satisfy conditions (4.20), i.e. will not be conjugate. However, if the initial point x_0 is in a close neighbourhood of the minimum x_* of a smooth convex function $f(x)$, then at any point of this region matrix $f''(x)$ is close enough to matrix $f''(x_*)$, i.e. the quadratic function

$$\varphi(x) = \frac{1}{2} (f''(x_*) (x - x_*), x - x_*) + f(x_*)$$

is a good approximation to the function $f(x)$. Thus, we can expect that the properties of vectors $p_0, \dots, p_k$ determined by methods of Sec. 4 will be close enough to the properties of conjugate vectors ($f''(x_*)$ -orthogonal) and therefore the properties of process (5.1) in which parameter α_k is chosen on condition that the minimum of function $f(x)$ occurs in the direction of p_k will be close enough to the properties of methods of conjugate directions. In other words, we can expect the methods of the preceding section prove sufficiently effective in minimizing nonquadratic functions too. In this case, the methods will no more yield the result after a finite number of steps since the conditions

$$(f''(x_*) p_k, p_i) = 0, \quad i \neq k$$

will not be strictly satisfied with any initial point x_0 .

Iterative processes of the type (5.1) in which vector p_k is constructed by algorithms of Sec. 4 and the value of parameter α_k is chosen on condition that

$$f(x_k + \alpha_k p_k) = \min_{\alpha} f(x_k + \alpha p_k)$$

will be called as before methods of conjugate directions.

Note that the condition under which parameter α_k is chosen can be written also in the following form:

$$(f'_{k+1}, p_k) = (f'(x_k + \alpha_k p_k), p_k) = 0. \quad (5.2)$$

The object of this section is to substantiate the convergence of methods of conjugate directions in the minimization of nonquadratic functions and to obtain bounds on the rate of convergence.

Theorem on Convergence of the Methods

In what follows we shall assume that $f(x)$ is a strongly convex differentiable function whose first and second derivatives are continuous, i.e. that conditions

$$m \|y\|^2 \leq (f''(x) y, y) \leq M \|y\|^2, \quad m > 0 \quad (5.3)$$

are satisfied for all $x, y \in E^n$, and that a symmetric, strictly positive definite matrix has been chosen as H_0 , i.e.

$$m_0 \|y\|^2 \leq (H_0 y, y) \leq M_0 \|y\|^2, \quad m_0 > 0 \quad (5.4)$$

for all $y \in E^n$.

Processes of type (5.1) can be realized either with restoration of matrix H_k after a finite number of steps, or without such a reinitialization. Speaking of processes with restoration, say, after n steps we mean that with any $\xi = 0, 1, \dots$ matrix $H_{\xi n}$ is restored, i.e. $H_{\xi n} = H_0$.

From the beginning note the following fact. If a process with restoration of matrix H_k after a finite number of steps is being realized, then for any of the methods of conjugate directions the following condition is fulfilled:

$$\lim_{k \rightarrow \infty} \|f'(x_k)\| = 0 \quad (5.5)$$

since each first step of the process after restoration is a step of gradient descent, for which according to (5.3) the conditions of convergence of gradient methods (theorem 1.6) are satisfied and in the following steps between restorations we have a descent to the minimum of the function in the direction of motion. The fulfillment of condition (5.5) for a strictly convex function means that *any of the methods discussed in Sec. 4, if realized with restoration of matrix H_k after a finite number of steps, converges to the solution x_** . Therefore in order to judge the effectiveness of such a process it is important to obtain bounds on its rate of convergence.

Note that condition (5.5) for processes with restoration is satisfied not only for strictly convex functions but for any function if the fulfillment of (5.5) is guaranteed for it in applying gradient methods (see theorem 1.4).

However, if processes are realized without restoration of H_k , then their convergence must be substantiated. Besides, it is also necessary to estimate their rate of convergence.

Let us now formulate the theorem whose contents are the main result of this section.

Theorem 5.1. *For the minimization of function $f(x)$ which satisfies conditions (5.3) let there be applied process (5.1) in which the construction of matrix H_k is performed by one of the methods of Sec. 4 ((4.48), (4.49), (4.52)-(4.54), (4.69)-(4.71)) with restoration of H_k after n steps. If the value of α_k is chosen under the condition that the minimum of the function be in the direction of p_k , then the sequence $\{x_{\xi_n}\}$ whatever the initial point x_0 chosen converges to the solution at a superlinear rate.*

Let us outline the general scheme of the proof of this theorem. Suppose that the hypothesis is not true, i.e. that for the iterative processes described the following condition is satisfied with any k :

$$\|x_{k+1} - x_*\| \geq \lambda \|x_k - x_*\| \quad (5.6)$$

where $\lambda > 0$ is a constant. Using inequality (1.12) and the expression

$$\|f'(x)\| = \|f'(x) - f'(x_*)\| \leq M \|x - x_*\| \quad (5.7)$$

which hold for a function which satisfies condition (5.3), we find that condition (5.6) is equivalent to the following one:

$$\|f'_{k+1}\| \geq \delta \|f'_k\| \quad (5.8)$$

where $\delta > 0$ is a constant.

Studying the properties of process (5.1) and assuming that condition (5.8) is fulfilled we find that independently of what algorithm is used for constructing matrix H_k the following estimates hold:

$$C \|f'_k\| \leq \|r_k\| \leq N \|f'_k\| \quad (5.9)$$

where C, N are constants independent of k , $C > 0$ and

$$(e_{\xi n+i}, r_{\xi n+j}) = O(\|e_{\xi n+i}\| \|r_{\xi n+j}\|), \quad i \neq j, \quad 0 \leq i, j \leq n-1. \quad (5.10)$$

It will be demonstrated below that if these estimates are fulfilled, then sequence (5.1) converges to the solution at a superlinear rate. However, this contradicts our initial assumption (5.6) (or (5.8)), i.e. condition (5.6) cannot be satisfied for process (5.1). Using this fact it will be easy to establish that the theorem holds.

Thus, the pattern of the proof is the same for all the methods discussed, but the validity of (5.9) and (5.10) is established in different ways. The proof that these estimates hold for different algorithms will be given in the next subsection and here we shall describe that part of the proof, which all these methods have in common.

We shall make first a remark about the notations. In what follows for the simplicity we shall often in using vectors and parameters $r_{\xi n+i}, f'_{\xi n+i}, e_{\xi n+i}, \alpha_{\xi n+i}, \beta_{\xi n+i}, i = 0, 1, \dots, n-1$ omit index ξn , i.e. operate with vectors and parameters $r_i, f'_i, e_i, \alpha_i, \beta_i$ etc. However, it should be stressed that this is done only to simplify the written form; the real index of the corresponding quantity is $\xi n + i$.

We turn now to the proof of the theorem and assume that estimates (5.9) and (5.10) are satisfied. Using Lagrange's formula for operators we obtain

$$(e_i, r_j) = (f''_{ic} r_i, r_j) = (f''_i r_i, r_j) + ((f''_{ic} - f''_i) r_i, r_j). \quad (5.11)$$

where, as usual, index ic denotes an intermediate point in the corresponding segment:

$$x_{ic} = x_i + \theta r_i, \quad 0 \leq \theta \leq 1.$$

If $\|r_i\| \rightarrow 0$, then because of the uniform continuity of second derivatives of function $f(x)$ on set $S = \{x: f(x) \leq f(x_0)\}$ we have $\|f''_{ic} - f''_i\| \rightarrow 0$ and it follows from (5.11) that, if (5.10) is satisfied, estimates

$$(f''_i r_i, r_j) = o(\|r_i\| \|r_j\|) + o(\|e_i\| \|r_j\|), \quad i \neq j, \quad 0 \leq i, j \leq n-1$$

hold too.

Under conditions (5.3) $\|e_i\| = \|f'_{i+1} - f'_i\| \leq M \|r_i\|$, consequently $\|e_i\|$ and $\|r_i\|$ are of the same order of smallness. Taking this into account, we have

$$(f''_i r_i, r_j) = o(\|r_i\| \|r_j\|), \quad i \neq j, \quad 0 \leq i, \quad j \leq n-1. \quad (5.12)$$

If estimates (5.12) are fulfilled, then there are vectors

$$\bar{r}_i = r_i + \omega_i, \quad i = 0, \quad 1, \dots, n-1, \quad (5.13)$$

where $\|\omega_i\| = o(\|r_i\|)$, such that

$$(f''_{\xi n} \bar{r}_i, r_j) = 0, \quad i \neq j, \quad 0 \leq i, j \leq n-1. \quad (5.14)$$

This can be shown as follows.

Let us normalize vectors r_i :

$$\tilde{r}_i = \frac{r_i}{(f''_{\xi n} r_i, r_i)^{1/2}}.$$

Then $(f''_{\xi n} \tilde{r}_i, \tilde{r}_i) = 1$ and as $\xi \rightarrow \infty$ since process (5.1) is convergent (with restoration of H_k) and by (5.3) and (5.12) we have

$$(f''_{\xi n} \tilde{r}_i, \tilde{r}_j) \leq \frac{1}{m \|r_i\| \|r_j\|} [(f''_i r_i, r_j) + ((f''_{\xi n} - f''_i) r_i, r_j)] \rightarrow 0$$

$$i \neq j, \quad 0 \leq i, j \leq n-1.$$

Therefore if $\tilde{R}_{\xi n}$ is a matrix whose columns are vectors $\tilde{r}_i$ and $F_{\xi n} = \tilde{R}_{\xi n}^* f''_{\xi n} \tilde{R}_{\xi n}$, then as $\xi \rightarrow \infty$

$$F_{\xi n} \rightarrow I.$$

Since $F_{\xi n}^{-1} \tilde{R}_{\xi n}^* f''_{\xi n} \tilde{R}_{\xi n} = I$ we obtain, denoting $F_{\xi n}^{-1} \tilde{R}_{\xi n}^*$ by $Q_{\xi n}^*$,

$$Q_{\xi n}^* f''_{\xi n} \tilde{R}_{\xi n} = I. \quad (5.15)$$

Now, since $F_{\xi n} \rightarrow I$, we have also $F_{\xi n}^{-1} \rightarrow I$ and, consequently, $\tilde{R}_{\xi n}^* \rightarrow Q_{\xi n}^*$, i.e. vector-columns q_i of matrix $Q_{\xi n}^*$ can be written in the form

$$q_i = \tilde{r}_i + \tilde{\omega}_i, \quad i = 0, \quad 1, \dots, n-1$$

where $\|\tilde{\omega}_i\| \rightarrow 0$ as $\xi \rightarrow \infty$. Let us write the equalities obtained in the following form:

$$(f''_{\xi n} r_i, r_i)^{1/2} q_i = r_i + (f''_{\xi n} r_i, r_i)^{1/2} \tilde{\omega}_i.$$

Because of (5.15), vectors r_i and $\bar{r}_i = (f''_{\xi n} r_i, r_i)^{1/2} q_i$, $i = 0, 1, \dots, n-1$ satisfy conditions (5.14). At the same time, vectors $\bar{r}_i$ satisfy conditions (5.13) since by (5.3)

$$\frac{\|\omega_i\|}{\|r_i\|} = \frac{(f''_{\xi n} r_i, r_i)^{1/2} \|\tilde{\omega}_i\|}{\|r_i\|} \leq \frac{M^{1/2} \|r_i\|}{\|r_i\|} \|\tilde{\omega}_i\| \rightarrow 0.$$

Thus it has been shown that (5.14) holds.

Vectors $\bar{r}_i$ with sufficiently large ξ are linearly independent. Indeed, let there be factors δ_i , $i = 0, 1, \dots, n-1$ (of which at least two are nonzero) such that $\sum_{i=0}^{n-1} \delta_i \bar{r}_i = 0$. If $\delta_0 \neq 0$, then we obtain

$$\delta_0 (f''_{\xi n} r_0, \bar{r}_0) + \sum_{j=1}^n \delta_j (f''_{\xi n} r_0, \bar{r}_j) = 0.$$

However, with sufficiently large values of ξ this equality is not satisfied. Indeed, since $\|\omega_i\| = o(\|r_i\|)$ and as $\xi \rightarrow \infty$, $\|r_i\| \rightarrow 0$, with sufficiently large ξ according to (5.3), we have

$$(f''_{\xi n} r_0, \bar{r}_0) = (f''_{\xi n} r_0, r_0) + (f''_{\xi n} r_0, \omega_0) \neq 0,$$

and at the same time $(f''_{\xi n} r_0, \bar{r}_j) = 0$, $j = 1, \dots, n-1$, because of (5.14). Thus we come to a contradiction, i.e. vectors $\bar{r}_i$, $i = 0, 1, \dots, n-1$ are really linearly independent.

Let $z_{\xi n}$ be the minimum of the quadratic function

$$\varphi(x) = (f'_{\xi n}, x - x_{\xi n}) + \frac{1}{2} (f''_{\xi n} (x - x_{\xi n}), x - x_{\xi n}).$$

Let us write vector $z_{\xi n} - x_{\xi n}$ in the form

$$z_{\xi n} - x_{\xi n} = \sum_{i=0}^{n-1} a_i \bar{r}_i. \quad (5.16)$$

Since $\varphi'(z_{\xi n}) = f'_{\xi n} + f''_{\xi n} (z_{\xi n} - x_{\xi n}) = 0$, then using (5.16) we obtain

$$\sum_{i=0}^{n-1} a_i f''_{\xi n} \bar{r}_i = -f'_{\xi n}.$$

Hence, taking into account (5.14), it follows that coefficients a_i can be calculated by the following formulas:

$$a_i = -\frac{(f'_{\xi n}, r_i)}{(f''_{\xi n} \bar{r}_i, r_i)}, \quad i = 0, 1, \dots, n-1.$$

Let us write the numerator on the right-hand side in the following form:

$$(f'_{\xi n}, r_i) = (f'_0 - f'_1 + f'_1 - \dots - f'_{i+1} + f'_{i+1}, r_i) = - \sum_{j=0}^i (e_j, r_i)$$

(we have taken into account that by (5.2) $(f'_{i+1}, r_i) = 0$). Hence, having in mind estimates (5.10), it follows that

$$(f'_{\xi n}, r_i) = -(e_i, r_i) + \sum_{j=0}^{i-1} o(\|r_i\| \|e_j\|). \quad (5.17)$$

According to conditions (5.8) and (5.9) all of the vectors $r_0, \dots, r_{n-1}$ are of the same order of smallness (recall that vectors $r_{\xi n+i}$ are practically meant). Since as was mentioned above $\|e_i\| \leq M\|r_i\|$, vectors $e_0, \dots, e_{n-1}$ are of the same order of smallness. Taking into account the above remarks the equalities (5.17) can be written in the following form:

$$(f'_{\xi n}, r_i) = -(e_i, r_i) + o(\|r_i\|^2), \quad i = 0, 1, \dots, n-1.$$

Further, taking into account (5.13), we find that

$$\begin{aligned} (f''_{\xi n} \bar{r}_i, r_i) &= (f''_{\xi n} r_i, r_i) + (f''_{\xi n} \omega_i, r_i) \\ &= (f''_{ic} r_i, r_i) + ((f''_{\xi n} - f''_{ic}) r_i, r_i) + (f''_{\xi n} \omega_i, r_i) = (e_i, r_i) + o_1(\|r_i\|^2). \end{aligned}$$

Thus

$$a_i = \frac{(e_i, r_i) + o(\|r_i\|^2)}{(e_i, r_i) + o_1(\|r_i\|^2)}.$$

By (5.3), we have

$$(e_i, r_i) = (f''_{ic} r_i, r_i) \geq m \|r_i\|^2, \quad i = 0, 1, \dots, n-1. \quad (5.18)$$

Consequently, as $\xi \rightarrow \infty$ (i.e. as $\|r_i\| \rightarrow 0$)

$$a_i \rightarrow 1, \quad i = 0, 1, \dots, n-1. \quad (5.19)$$

Since $x_{(\xi+1)n} - x_{\xi n} = \sum_{i=0}^{n-1} r_i$, we have

$$x_{(\xi+1)n} - z_{\xi n} = (x_{(\xi+1)n} - x_{\xi n}) - (z_{\xi n} - x_{\xi n}) = \sum_{i=0}^{n-1} (r_i - a_i \bar{r}_i).$$

Hence, taking into account (5.13) and (5.19), we obtain

$$\|x_{(\xi+1)n} - z_{\xi n}\| = \sum_i o(\|r_i\|)$$

or using (5.8) and (5.9)

$$\|x_{(\xi+1)n} - z_{\xi n}\| = o(\|f'_{\xi n}\|). \quad (5.20)$$

Since $z_{\xi n} - x_{\xi n} = -(f''_{\xi n})^{-1}f'_{\xi n}$ and taking into account (5.20), we have

$$\begin{aligned} x_{(\xi+1)n} - x_{\xi n} &= (x_{(\xi+1)n} - z_{\xi n}) + (z_{\xi n} - x_{\xi n}) \\ &= -(f''_{\xi n})^{-1}f'_{\xi n} + \eta_{\xi n} \end{aligned}$$

where $\|\eta_{\xi n}\| = o(\|f'_{\xi n}\|)$.

It follows that there is a sequence of matrices $D_{\xi n}^{-1} \rightarrow (f''_{\xi n})^{-1}$ such that

$$x_{(\xi+1)n} - x_{\xi n} = -D_{\xi n}^{-1}f'_{\xi n}. \quad (5.21)$$

(We can take, for instance, that

$$D_{\xi n}^{-1} = (f''_{\xi n})^{-1} + \frac{z_{\xi n} - x_{(\xi+1)n}}{(f'_{\xi n}, f'_{\xi n})} (f'_{\xi n})^*.)$$

Equality (5.21) shows that sequence $\{x_{\xi n}\}$, $\xi = 0, 1, \dots$ converges at a superlinear rate to the solution; the corresponding bounds on the rate of convergence can be obtained just as it was done in theorem 3.1 for sequence $\{x_k\}$.

Thus assuming that condition (5.6) is satisfied and taking estimates (5.10) and (5.9) to be satisfied, we have demonstrated that for $\{x_{\xi n}\}$ inequality

$$\|x_{(\xi+1)n} - x_*\| \leq \lambda_{\xi n} \|x_{\xi n} - x_*\| \quad (5.22)$$

where $\lambda_{\xi n} \rightarrow 0$ as $\xi \rightarrow \infty$, holds. However, if condition (5.6) holds, then inequality (5.22) cannot be satisfied since with (5.6) fulfilled we have

$$\|x_{(\xi+1)n} - x_*\| \geq \lambda^n \|x_{\xi n} - x_*\|. \quad (5.23)$$

Thus we have come to a contradiction. This means that condition (5.6) (or (5.23)) cannot be fulfilled for process (5.1).

The impossibility of fulfilling condition (5.6) with any k means, in fact, that for process (5.1) (with reinitialization) inequality (5.23) cannot be fulfilled in any subsequence $\{\xi_m\}$, $m = 0, 1, \dots$. If there were a sequence $\{\xi_m\}$ such that

$$\|x_{(\xi_m+1)n} - x_*\| \geq \lambda^n \|x_{\xi_m n} - x_*\|, \quad (5.24)$$

then with any $\xi_m n \leq k \leq (\xi_m + 1)n$ for methods with reinitialization estimates (5.9), (5.10) would be fulfilled; this will be found in studying the properties of such processes in the next subsection. Therefore, repeating the above argument we would have concluded that at iterations which correspond to sequence $\{\xi_m\}$ inequality (5.22) is fulfilled and this contradicts (5.24).

Thus, for process (5.1) with restoration of matrix H_k inequality (5.24) cannot be fulfilled. It follows that for any constant $\lambda > 0$ there is a number T such that with $\xi > T$ condition (5.22) is fulfilled, i.e. sequence $\{x_{\xi n}\}$ converges to the solution at a superlinear rate.

Study of Properties of Different Algorithms

We turn now to the proof of the validity of estimates (5.9), (5.10) for different methods of conjugate directions with restoration of matrix H_k after n steps, assuming that inequality (5.6) (or (5.8)) is fulfilled.

The fact that for any of these methods the estimates hold is established by induction; it is demonstrated that estimates (5.9), (5.10) take place with $i \neq j$, $i, j = 0, 1$; and then supposing that these estimates take place with $0 \leq i, j \leq \tau < n - 1$ we prove that they remain valid also with $0 \leq i, j \leq \tau + 1$.

1. Method (4.48). If restoration of matrix (4.48) is performed after a finite number of steps, then with any k matrix H_k is bounded:

$$\|H_k\| \leq L, L < \infty. \quad (5.25)$$

We show now how this can be proved.

By (5.2),

$$(H_k f'_k, f'_{k+1}) = - (p_k, f'_{k+1}) = 0.$$

Therefore

$$(H_k e_k, e_k) = (H_k f'_k, f'_k) + (H_k f'_{k+1}, f'_{k+1}). \quad (5.26)$$

Since H_k is positive definite (Sec. 4), we have $(H_k e_k, e_k) \geq - (p_k, f'_k) = (p_k, e_k)$. Hence, according to (5.18)

$$(H_k e_k, e_k) \geq \frac{m}{\alpha_k} \|r_k\|^2. \quad (5.27)$$

Taking into account moreover that $\|e_k\| \leq M \|r_k\|$, we obtain from (4.48) that

$$\|H_{k+1}\| \leq \|H_k\| + \frac{\|r_k\|^2}{m \|r_k\|^2} + \frac{\alpha_k \|H_k\|^2 M^2 \|r_k\|^2}{m \|r_k\|^2}.$$

Using condition (5.4) one can easily ascertain that $\alpha_{\xi n} \leq \bar{\alpha} < \infty$ and on the strength of this it follows from the recursive inequality for $\|H_{k+1}\|$ that estimate (5.25) holds for $H_{\xi n+1}$. On this ground we shall prove below by induction that $\alpha_{\xi n+i} \leq \bar{\alpha} < \infty$ with any $i = 1, \dots, n - 1$. Taking this into account we find that (5.25) holds.

Let us prove now that with $i = 1$ the following relations hold:

$$\begin{aligned} (r_1, e_0) &= 0, & (e_1, r_0) &= o(\|r_0\| \|r_1\|), \\ C_1 \|f'_1\| &\leq \|r_1\| \leq N_1 \|f'_1\| \end{aligned} \quad (5.28)$$

where the constants N_1, C_1 are independent of k and $C_1 > 0$. The first of these estimates is found as follows:

$$(r_1, e_0) = -\alpha_1 (H_1 f'_1, e_0) = -\alpha_1 (f'_1, H_1 e_0).$$

But $H_1 e_0 = r_0$, therefore $(r_1, e_0) = -\alpha_1(f'_1, r_0) = 0$. Further,
 $(e_1, r_0) = (f''_{1c} r_1, r_0) = (r_1, f''_{0c} r_0)$
 $+ (r_1, (f''_{1c} - f''_{0c}) r_0) = (r_1, e_0) + o(\|r_1\| \|r_0\|) = o(\|r_1\| \|r_0\|).$

Let us now show that the estimates hold for $\|r_1\|$. It follows from (4.48), taking into account (5.2) and (5.26), that

$$(H_1 f'_1, f'_1) = (H_0 f'_1, f'_1) - \frac{(H_0 e_0, f'_1)^2}{(H_0 e_0, e_0)} = (H_0 f'_1, f'_1) \frac{1}{1 + \frac{(H_0 f'_1, f'_1)}{(H_0 f'_0, f'_0)}}.$$

Using estimates (1.14), (1.15) and (5.7) we deduce that for a function which satisfies (5.3)

$$m \left(1 + \frac{m}{M}\right) (f(x) - f_*) \leq \|f'(x)\|^2 \leq \frac{2M^2}{m} (f(x) - f_*). \quad (5.29)$$

Taking into account estimates (5.4) and (5.29) we have on set $S_0 = \{x: f(x) \leq f(x_0)\}$

$$\frac{(H_0 f'_1, f'_1)}{(H_0 f'_0, f'_0)} \leq \frac{M_0 \|f'_1\|^2}{m_0 \|f'_0\|^2} \leq \frac{d_1 (f_1 - f_*)}{d_2 (f_0 - f_*)} \leq \frac{d_1}{d_2}$$

where d_1, d_2 are constants independent of ξ . By virtue of this,

$$(H_1 f'_1, f'_1) \geq \frac{(H_0 f'_1, f'_1)}{1 + \frac{d_1}{d_2}} \geq a_1 \|f'_1\|^2. \quad (5.30)$$

where $a_1 = m_0 / \left(1 + \frac{d_1}{d_2}\right)$ is independent of ξ .

Let us use now inequality (5.30) in order to estimate the value of parameter $\alpha_{\xi_{n+1}}$. Since

$$f_2 - f_1 = \alpha_1 (f'_1, p_1) + \frac{\alpha_1^2}{2} (f''_{1c} p_1, p_1)$$

and α_1 is chosen according to condition (5.2), it is clear that

$$-\frac{(f'_1, p_1)}{M \|p_1\|^2} \leq \alpha_1 \leq -\frac{(f'_1, p_1)}{m \|p_1\|^2}.$$

Now by (5.30), we have $-(f'_1, p_1) = (H_1 f'_1, f'_1) \geq a_1 \|f'_1\|^2$ and by (5.25), $\|p_1\| = \|H_1 f'_1\| \leq L \|f'_1\|$; taking these estimates into account, we have $\alpha_1 \geq \frac{a_1}{ML^2} = \alpha > 0$. At the same time it follows from (5.30) that $\|p_1\| \geq a_1 \|f'_1\|$. Using this estimate we can easily establish that $\alpha_1 \leq \frac{L}{ma_1^2} = \bar{\alpha} < \infty$. Thus we find that

$$N_1 \|f'_1\| = \bar{\alpha} L \|f'_1\| \geq \|r_1\| = \alpha_1 \|H_1^* f'_1\| \geq \alpha a_1 \|f'_1\| = C_1 \|f'_1\|$$

where constants N_1, C_1 are independent of ξ , i.e. we have established that estimates (5.28) hold.

Suppose that the estimates

$$(r_i, e_j) = o(\|r_i\| \|r_j\|), \quad i \neq j, \quad 0 \leq i, \quad j \leq \tau < n - 1, \quad (5.31)$$

$$C_i \|f'_i\| \leq \|r_i\| \leq N_i \|f'_i\|, \quad 0 \leq i \leq \tau \quad (5.32)$$

where constants $N_i, C_i > 0$ are independent of ξ , hold. Let us show that similar estimates take place also with $0 \leq i, j \leq \tau + 1$

$$(f'_{\tau+1}, r_j) = (f'_{j+1}, r_j) + (e_{j+1} + \dots + e_\tau, r_j), \quad 0 \leq j < \tau. \quad (5.33)$$

According to condition (5.8) and estimates (5.32) quantities $\|f'_{\tau+1}\|, \|f'_\tau\|$ and $\|r_i\|$ with any $0 \leq i \leq \tau$ are of the same order of smallness. Taking this into account and using conditions (5.2) and (5.31) we find according to (5.33) that

$$(f'_{\tau+1}, r_j) = o(\|f'_{\tau+1}\| \|r_j\|) = o(\|r_j\|^2), \quad 0 \leq j < \tau.$$

Since $(f'_{\tau+1}, r_\tau) = 0$, according to (5.2), we obtain finally

$$(f'_{\tau+1}, r_j) = o(\|r_j\|^2), \quad 0 \leq j \leq \tau. \quad (5.34)$$

Let us estimate now the quantity $(H_{\tau+1}f'_{\tau+1}, f'_{\tau+1})$. Using formula (4.48) and taking into account (5.26) we obtain for any $0 \leq j \leq \tau$:

$$\begin{aligned} (H_{j+1}f'_{\tau+1}, f'_{\tau+1}) &= (H_j f'_{\tau+1}, f'_{\tau+1}) + \frac{(r_j, f'_{\tau+1})^2}{(r_j, e_j)^2} \\ &- \frac{(H_j e_j, f'_{\tau+1})^2}{(H_j e_j, e_j)} \geq \frac{1}{(H_j e_j, e_j)} [(H_j f'_{\tau+1}, f'_{\tau+1})(H_j f'_{j+1}, f'_{j+1}) \\ &+ (H_j f'_{\tau+1}, f'_{\tau+1})(H_j f'_j, f'_j) - (H_j f'_{j+1}, f'_{\tau+1})^2 \\ &- (H_j f'_j, f'_{\tau+1})^2 + 2(H_j f'_{j+1}, f'_{\tau+1})(H_j f'_j, f'_{\tau+1})]. \end{aligned}$$

On the right-hand side of this inequality the difference between the first and the third terms of the numerator, by Cauchy-Buniakowski's inequality, is nonnegative. Taking into account estimates (5.34), (5.27) and (5.25) and that $\alpha_j, j \leq \tau$ is bounded, it is easy to ascertain that the ratio of the last two terms of the numerator to the denominator is of the order of $o(\|r_j\| \|f'_{\tau+1}\|) = o(\|f'_{\tau+1}\|^2)$. Hence

$$(H_{j+1}f'_{\tau+1}, f'_{\tau+1}) \geq \frac{(H_j f'_{\tau+1}, f'_{\tau+1})(H_j f'_j, f'_j)}{(H_j e_j, e_j)} - o(\|f'_{\tau+1}\|^2).$$

Estimates (5.32) imply that there are constants a_i independent of ξ such that $(H_j f'_j, f'_j) = -(p_j, f'_j) \geq a_j \|f'_j\|^2$. Making use of this fact and of (5.25) we have

$$\begin{aligned} (H_{j+1}f'_{\tau+1}, f'_{\tau+1}) &\geq \frac{a_j \|f'_j\|^2}{ML \|r_j\|^2} (H_j f'_{\tau+1}, f'_{\tau+1}) \\ &- o(\|f'_{\tau+1}\|^2) \geq \tilde{a}_j (H_j f'_{\tau+1}, f'_{\tau+1}) - o(\|f'_{\tau+1}\|^2) \end{aligned} \quad (5.35)$$

where $\tilde{a}_j > 0$ and is independent of ξ (by (5.32)).

It was noted in the preceding subsection that for processes with restoration of H_k as $k \rightarrow \infty$ we have $\|f'_k\| \rightarrow 0$. Therefore, it follows from inequalities (5.35), taking into account that matrix H_k is positive definite, that if with any ξ we have $(H_j f'_{\tau+1}, f'_{\tau+1}) \geq \gamma_j \|f'_{\tau+1}\|^2$ where $\gamma_j > 0$ and is independent of ξ , then there is a constant $\gamma_{j+1} > 0$ such that with any ξ we shall have $(H_{j+1} f'_{\tau+1}, f'_{\tau+1}) \geq \gamma_{j+1} \|f'_{\tau+1}\|^2$. But in estimating the quantity $(H_1 f'_{\tau+1}, f'_{\tau+1})$ we find, since $(H_0 f'_{\tau+1}, f'_{\tau+1}) \geq m_0 \|f'_{\tau+1}\|^2$, that there is a constant γ_1 such that $(H_1 f'_{\tau+1}, f'_{\tau+1}) \geq \gamma_1 \|f'_{\tau+1}\|^2$ with any ξ . Taking this into account, our argument by induction shows that there is a constant $a_{\tau+1}$ independent of ξ and such that $(H_{\tau+1} f'_{\tau+1}, f'_{\tau+1}) \geq a_{\tau+1} \|f'_{\tau+1}\|^2$. We establish now just as we did above that

$$\frac{a_{\tau+1}}{ML^2} \leq -\frac{(f'_{\tau+1}, p_{\tau+1})}{M \|p_{\tau+1}\|^2} \leq \alpha_{\tau+1} \leq -\frac{(f'_{\tau+1}, p_{\tau+1})}{m \|p_{\tau+1}\|^2} \leq \frac{L}{ma_{\tau+1}^2}.$$

Therefore, we have

$$N_{\tau+1} \|f'_{\tau+1}\| \geq \|r_{\tau+1}\| = \alpha_{\tau+1} \|H_{\tau+1} f'_{\tau+1}\| \geq C_{\tau+1} \|f'_{\tau+1}\|. \quad (5.36)$$

Let us show now that

$$H_{\tau+1} e_j = r_j + \eta_j, \quad 0 \leq j \leq \tau \quad (5.37)$$

where $\|\eta_j\| = o(\|r_j\|)$.

Multiplying both sides of formula (4.48) by e_j we obtain

$$H_{s+1} e_j = H_s e_j + \frac{r_s (r_s, e_j)}{(r_s, e_s)} - \frac{(H_s e_s, e_j) H_s e_s}{(H_s e_s, e_s)}. \quad (5.38)$$

If we assume that with a certain s , $j+1 \leq s \leq \tau$, equalities $H_s e_j = r_j + \tilde{\eta}_j$ take place where $\|\tilde{\eta}_j\| = o(\|r_j\|)$, then using estimates (5.34), (5.27), (5.25) and taking into account that all of the quantities $\|r_s\|$ are of the same order of smallness we also have by (5.38) that $H_{s+1} e_j = r_j + \eta_j$, where $\|\eta_j\| = o(\|r_j\|)$. But $H_{j+1} e_j = r_j$, and we establish by induction that equalities (5.37) hold true.

Taking into account (5.37) we have

$$(r_{\tau+1}, e_j) = -\alpha_{\tau+1} (H_{\tau+1} f'_{\tau+1}, e_j) = -\alpha_{\tau+1} (f'_{\tau+1}, r_j + \eta_j),$$

therefore by (5.34), we find that

$$(r_{\tau+1}, e_j) = o(\|r_j\|^2) + o(\|f'_{\tau+1}\| \|r_j\|), \quad 0 \leq j \leq \tau.$$

Inequalities (5.8) and (5.36) show that $\|r_{\tau+1}\|$ is of the same order of smallness as $\|f'_{\tau+1}\|$ and consequently as $\|r_j\|$, $0 \leq j \leq \tau$. It follows that

$$(r_{\tau+1}, e_j) = o(\|r_{\tau+1}\| \|r_j\|) = o(\|r_{\tau+1}\|^2), \quad 0 \leq j \leq \tau. \quad (5.39)$$

Taking this into account we establish in a manner analogous to that used before with $i = 1$ that also

$$(e_{\tau+1}, r_j) = o(\|r_{\tau+1}\|^2), \quad 0 \leq j \leq \tau. \quad (5.40)$$

The relations (5.36), (5.39) and (5.40) show that estimates (5.31) and (5.32) really take place with $\tau + 1$ too.

Thus it has been established that estimates (5.9), (5.10) hold for method (4.48) if it is assumed that process (5.1) is realized with restoration of matrix H_k after a finite number of steps.

The above argument can be repeated word for word if we assume that inequality (5.24) (or the corresponding inequality $\|f'_{(\xi_{m+1})n}\| \geq \geq \delta^n \|f'_{\xi_m n}\|$) is satisfied rather than condition (5.6) (or (5.23)) and consider only iterations which correspond to subsequence $\{\xi_m\}$. At these iterations estimates (5.9) and (5.10) remain valid.

The superlinear rate of convergence of the method follows from this fact as was shown in the preceding subsection.

2. Method (4.49). If matrix H_k is restored after a finite number of steps, then with any k the matrix has a bound. This follows from inequality

$$\|H_{k+1}\| \leq \|H_k\| + \frac{\|r_k\|^2}{m\|r_k\|^2} + \frac{\|H_k\| M \|r_k\|^2}{m\|r_k\|^2}.$$

With $i = 1$

$$(H_1^* f'_1, f'_1) = (H_0 f'_1, f'_1) \geq m_0 \|f'_1\|^2.$$

Making use of these relations and reasoning as in studying method (4.48) we establish that estimates (5.28) hold and then assuming that estimates (5.31) and (5.32) hold we demonstrate that estimate (5.34) holds.

Further we have that

$$(H_{\tau+1}^* f'_{\tau+1}, f'_{\tau+1}) = (H_0^* f'_{\tau+1}, f'_{\tau+1}) + \sum_{i=0}^{\tau} \frac{(f'_{\tau+1}, r_i)(r_i - H_i e_i, f'_{\tau+1})}{(r_i, e_i)}.$$

Using estimates (5.18), (5.34) and the fact that H_k has a bound and taking into account that all the quantities $\|r_i\|$, $\|e_i\|$, $\|f'_{\tau+1}\|$ are of the same order of smallness we find that

$$(H_{\tau+1}^* f'_{\tau+1}, f'_{\tau+1}) \geq m_0 \|f'_{\tau+1}\|^2 + o(\|f'_{\tau+1}\|^2).$$

Consequently with sufficiently small values of $\|f'_{\tau+1}\|$ (i.e. with sufficiently large ξ) we have

$$(H_{\tau+1}^* f'_{\tau+1}, f'_{\tau+1}) \geq a_{\tau+1} \|f'_{\tau+1}\|^2$$

where $a_{\tau+1} > 0$ and is independent of ξ . This being so, we find that $\bar{\alpha} \geq \alpha_{\tau+1} \geq \alpha > 0$ and $C_{\tau+1} \|f'_{\tau+1}\| \leq \|r_{\tau+1}\| \leq N_{\tau+1} \|f'_{\tau+1}\|$.

Using equalities

$$H_{s+1} e_i = H_s e_i + (r_s - H_s e_s) \frac{(r_s, e_i)}{(r_s, e_s)}, \quad i + 1 \leq s \leq \tau,$$

taking into account estimates (5.31), (5.18), the fact that H_k has a bound and reasoning as we did in studying (4.48), we ascertain that matrix $H_{\tau+1}$ satisfies equations (5.37) and consequently estimates (5.39) and (5.40) remain valid.

This completes the proof that our reasoning by induction holds.

The study of method (4.52) is carried out in a similar way.

Remark. If at an iteration of the initial stage of the process we find that $p_i = -H_i^* f'_i = 0$ then it is necessary to restart the process restoring matrix H_0 .

3. Method (4.53). The technique of the proofs pertaining to matrix H_k in this case is just the same as in method (4.49). Note only that in this method matrix $H_{\tau+1}$ satisfies equations

$$H_{\tau+1} e_j = \eta_j, \quad 0 \leq j \leq \tau$$

where $\|\eta_j\| = o(\|r_j\|)$, rather than conditions (5.37). Though this simplifies the obtaining of estimate (5.39).

The study of method (4.54) is analogous.

4. Method (4.69). Matrix H_k (4.69) determines vector p_k (4.63) in which $\tilde{\beta}_k$ is calculated by formula (4.64).

For this method

$$\|H_k\| \leq M_0 + \frac{M_0^2 \|f'_k\|^2 + M_0 \|f'_k\| \|H_{k-1}\| \|f'_{k-1}\|}{m_0 \|f'_k\|^2}.$$

By (5.8) with any k we have $\frac{\|f'_{k-1}\|}{\|f'_k\|} \leq \frac{1}{\delta} = d_0$, where d_0 is independent of ξ . Taking this into account we find that with any k for the process with restoration, $\|H_k\| \leq L$.

Let us show now that with any k $\tilde{\beta}_k \leq \bar{\beta} < 1$, i.e. that

$$1 - \tilde{\beta}_k \geq 1 - \bar{\beta} = \beta > 0 \quad (5.41)$$

where β is independent of ξ .

Using (4.65) we find that

$$\begin{aligned} \tilde{\beta}_k &= \frac{(H_0 f'_k, f'_k)}{(H_0 f'_k, f'_k) + (1 - \tilde{\beta}_{k-1}) (H_0 f'_{k-1}, f'_{k-1})} \\ &= \frac{1}{1 + (1 - \tilde{\beta}_{k-1}) \frac{(H_0 f'_{k-1}, f'_{k-1})}{(H_0 f'_k, f'_k)}}. \end{aligned}$$

Because of (5.4) and (5.8) with any k

$$\frac{(H_0 f'_{k-1}, f'_{k-1})}{(H_0 f'_k, f'_k)} \leq \frac{M_0 \|f'_{k-1}\|^2}{m_0 \|f'_k\|^2} \leq \gamma < \infty$$

where γ is independent of ξ .

Besides, using (4.65) and (4.64) it is easy to establish by induction that with any k we have $0 < \tilde{\beta}_k < 1$. Consequently

$$\tilde{\beta}_k \leq \frac{1}{1 + (1 - \tilde{\beta}_{k-1}) \gamma}.$$

Hence taking into account that $\tilde{\beta}_{\xi n} = 0$, $\xi = 0, 1, \dots$, it can be established that (5.41) holds. It follows from (4.65) using (5.41) that with any k

$$-(p_k, f'_k) \geq \beta m_0 \|f'_k\|^2,$$

i.e.

$$\|p_k\| \geq \beta m_0 \|f'_k\|.$$

Taking into account this inequality and the fact that H_k is bounded as in the methods considered above we establish that $0 < \alpha \leq \alpha_k \leq \bar{\alpha}$ and $C \|f'_k\| \leq \|r_k\| \leq N \|f'_k\|$. Thus it has been proved that estimates (5.9) hold.

Let us show that estimates (5.10) hold. With $k = \xi n + 1$ we have

$$(p_1, e_0) = -(H_0 f'_1, e_0) + \frac{(H_0 f'_1, f'_1)}{(H_0 f'_1, f'_1) - (p_0, f'_0)} [(H_0 f'_1, e_0) + (p_0, e_0)].$$

But $(H_0 f'_1, e_0) = (H_0 f'_1, f'_1)$ and because of (4.61) $(p_0, e_0) = -(p_0, f'_0)$. Making use of these equalities we find that $(p_1, e_0) = 0$.

If estimates (5.31) hold, then in the same way as it was done for method (4.48) it can be proved that estimates (5.34) hold.

Further using (4.63) we establish that

$$(p_{\tau+1}, e_j) = (\tilde{\beta}_{\tau+1} - 1) (H_0 f'_{\tau+1}, e_j) + \tilde{\beta}_{\tau+1} (p_{\tau}, e_j).$$

Let us estimate the quantity $(H_0 f'_{\tau+1}, e_j)$. It follows from (4.63) that

$$H_0 f'_j = \frac{1}{\tilde{\beta}_j - 1} (p_j - \tilde{\beta}_j p_{j-1}). \quad (5.42)$$

Taking this into account we have

$$H_0 e_j = \frac{1}{\tilde{\beta}_{j+1} - 1} (p_{j+1} - \tilde{\beta}_{j+1} p_j) - \frac{1}{\tilde{\beta}_j - 1} (p_j - \tilde{\beta}_j p_{j-1}).$$

Using this expression and taking into account estimates (5.34), (5.41) we establish that

$$(f'_{\tau+1}, H_0 e_j) = o(\|p_{\tau+1}\|^2) = o(\|r_{\tau+1}\|^2), \quad 0 \leq j < \tau.$$

Since by (5.31) we have also $(p_{\tau}, e_j) = o(\|r_{\tau+1}\|^2)$, $0 \leq j \leq \tau$, we find that

$$(p_{\tau+1}, e_j) = o(\|r_{\tau+1}\|^2), \quad 0 \leq j < \tau.$$

With $j = \tau$

$$(p_{\tau+1}, e_{\tau}) = (\tilde{\beta}_{\tau+1} - 1) (H_0 f'_{\tau+1}, e_{\tau}) + \tilde{\beta}_{\tau+1} (p_{\tau}, e_{\tau}).$$

Hence using (4.64) and (4.61), we obtain:

$$(p_{\tau+1}, e_{\tau}) = - \frac{(H_0 f'_{\tau+1}, f'_{\tau}) (p_{\tau}, f'_{\tau})}{(H_0 f'_{\tau+1}, f'_{\tau+1}) - (p_{\tau}, f'_{\tau})}.$$

Since $(H_0 f'_{\tau+1}, f'_{\tau+1}) > 0$, we have

$$(p_{\tau+1}, e_{\tau}) \leq (H_0 f'_{\tau+1}, f'_{\tau}).$$

If quantity $(H_0 f'_{\tau+1}, f'_{\tau})$ is estimated using expression (5.42) with $j = \tau$, then making use of (5.34) and (5.41) we find that

$$(H_0 f'_{\tau+1}, f'_{\tau}) = o(\|r_{\tau+1}\|^2)$$

(it has been taken into account that since (5.9) holds, the quantities $\|r_i\|$ and $\|f'_i\|$, $i = 0, 1, \dots, n-1$ are of the same order of smallness). Thus

$$(p_{\tau+1}, e_j) = o(\|r_{\tau+1}\|^2), \quad 0 \leq j \leq \tau,$$

i.e. for the method under consideration estimates (5.39) hold. Consequently estimates (5.40) hold too.

Thus we have established that for method (4.69) assuming that condition (5.6) is fulfilled, estimates (5.9) and (5.10) hold too.

5. Method (4.71). By (4.67) and (5.4) we have for this method

$$-(p_k, f'_k) \geq m_0 \|f'_k\|^2. \quad (5.43)$$

Taking this into account we have

$$\|H_k\| \leq \|H_0\| + \frac{\|H_0\| \|f'_k\| \|H_{k-1}^*\| \|f'_{k-1}\|}{m_0 \|f'_{k-1}\|^2}.$$

Since with any k the ratio $\|f'_k\| / \|f'_{k-1}\|$ on set S_0 is a bounded quantity and $\|H_0\| \leq M_0$, we have

$$\|H_k\| \leq M_0 + d \|H_{k-1}\|$$

where d is a constant. It follows that if the matrix is restored after a finite number of steps, then with any k matrix H_k has a bound: $\|H_k\| \leq L$. Making use of this fact and estimate (5.43) we ascertain that with any k , $\alpha \geq \alpha_k \geq \alpha > 0$ and $N \|f'_k\| \geq \|r_k\| \geq C \|f'_k\|$.

Consequently, for method (4.71) estimates (5.9) hold. Let us prove that estimates (5.10) hold:

$$(p_{k+1}, e_k) = -(H_0 f'_{k+1}, e_k) - \frac{(H_0 f'_{k+1}, f'_{k+1})}{(p_k, f'_k)} (p_k, e_k).$$

Hence, taking (4.61) into account,

$$(p_{k+1}, e_k) = -(H_0 f'_{k+1}, e_k) + (H_0 f'_{k+1}, f'_{k+1}) = (H_0 f'_{k+1}, f'_k). \quad (5.44)$$

With $k = \xi n$, (5.44) yields

$$(p_1, e_0) = (H_0 f'_1, f'_0) = (f'_1, p_0) = 0.$$

Using (4.66) we can obtain the following expressions:

$$\begin{aligned} (p_{\tau+1}, e_j) &= - (H_0 f'_{\tau+1}, e_j) + \beta_{\tau+1} (p_{\tau}, e_j), \\ H_0 f'_j &= - p_j + \beta_j p_{j-1}, \\ H_0 e_j &= - p_{j+1} + p_j - \beta_{j+1} p_j + \beta_j p_{j-1}. \end{aligned}$$

If we assume that estimates (5.31) hold, then arguing as in method (4.48) we can prove that estimates (5.34) hold. Note also that

$$\beta_k = - \frac{(H_0 f'_k, f'_k)}{(p_{k-1}, f'_{k-1})} \leq \frac{M_0 \|f'_k\|^2}{m_0 \|f'_{k-1}\|^2} \leq \frac{d_1}{d_2}.$$

Taking this into account and reasoning as we did in method (4.69) we establish that

$$(p_{\tau+1}, e_j) = o(\|r_{\tau+1}\|^2), \quad 0 \leq j < \tau$$

and, besides,

$$(H_0 f'_\tau, f'_{\tau+1}) = o(\|r_{\tau+1}\|^2)$$

hence it follows from (5.44) that

$$(p_{\tau+1}, e_\tau) = o(\|r_{\tau+1}\|^2).$$

Thus

$$(p_{\tau+1}, e_j) = o(\|r_{\tau+1}\|^2), \quad 0 \leq j \leq \tau$$

and this proves that estimates (5.39) hold. The validity of estimates (5.40) is established in the same manner as it was done for method (4.48). Consequently, estimates (5.10) hold for the method under consideration.

In studying method (4.48) we noted that the proof of estimates (5.9), (5.10) can be given assuming that condition (5.24) is fulfilled; it is only necessary to repeat the above argument for corresponding iterations. This remark is applicable to other methods considered above and this implies that they converge at a superlinear rate.

Further Study of the Rate of Convergence

1. Suppose now that matrix $f''(x)$ besides conditions (5.3) satisfies Lipschitz' condition (2.8). In this case it is possible to obtain a more precise bound on the rate of convergence of sequence $\{x_{\xi n}\}$.

To make referring more convenient, we shall give the different relations which hold if (5.3) does (many of them were often used before):

$$m \|x - x_*\| \leq \|f'(x)\| \leq M \|x - x_*\| \quad (5.45)$$

$$d_1 [f(x) - f(x_*)] \leq \|f'(x)\|^2 \leq d_2 [f(x) - f(x_*)] \quad (5.46)$$

(the constants d_1 and d_2 are independent of the choice of point x)

$$m \|r_k\| \leq \|e_k\| \leq M \|r_k\|. \quad (5.47)$$

Let x, y be arbitrary points

$$f(y) \leq f(x). \quad (5.48)$$

Then making use of (5.46) we establish that

$$\|f'(y)\| \leq C \|f'(x)\|. \quad (5.49)$$

Here and further on in this subsection C denotes various constants (not equal to zero) which are independent of the choice of points $x, y \in E^n$.

If (5.48) is satisfied, we have by (5.45) and (5.49)

$$\|y - x_*\| \leq C \|x - x_*\|. \quad (5.50)$$

2. Suppose that for the iterative processes being studied the following estimate holds:

$$\|f'_{\xi n+i}\| \|f'_{\xi n+i+1}\| \leq \lambda_{\xi} \|f'_{\xi n+j}\|, \quad 0 \leq i < j \leq n-1. \quad (5.51)$$

Here and further on, λ_{ξ} will denote different variables tending to zero as $\xi \rightarrow \infty$.

In what follows we shall limit ourselves to the study of the properties of method (4.48). However, the results obtained (lemma 5.1, theorem 5.2) hold also for other algorithms of conjugate directions.

Lemma 5.1. *Let process (5.1) be used for the minimization of the twice continuously differentiable function $f(x)$ which also satisfies conditions (5.3) and (2.8); in this process the construction of matrix H_k is performed by formula (4.48). Then if inequalities (5.51) hold, estimates (5.9) also hold and moreover*

$$\begin{aligned} |(r_{\xi n+i}, e_{\xi n+j})| &\leq C \|r_{\xi n+t}\|^2 \|r_{\xi n+t+1}\| \\ t &= \min \{i, j\}, \quad i \neq j, \quad 0 \leq i, \quad j \leq n-1. \end{aligned} \quad (5.52)$$

This lemma is proved in the same way as estimates (5.9) and (5.10) for method (4.48) in the subsection on p. 111; only the order of smallness of some quantities is determined more precisely. Therefore,

we abstain from giving a detailed proof and shall dwell only on the changes involved.

With $i = 1$

$$(e_1, r_0) = (r_1, e_0) + (r_1, (f''_{1c} - f''_{0c})r_0).$$

Taking (2.8) into account, we obtain

$$\begin{aligned} \|f''_{1c} - f''_{0c}\| &\equiv \|f''(x_1 + \theta_1 r_1) - f''(x_0 + \theta_0 r_0)\| \\ &\leq R(\|x_1 - x_0\| + \|r_1\| + \|r_0\|) \leq R(2\|r_0\| + \|r_1\|). \end{aligned}$$

By (5.50) and (5.45), we have

$$\|r_k\| \leq \|x_k - x_*\| + \|x_{k+1} - x_*\| \leq C\|f'_k\|. \quad (5.53)$$

Using (5.53) and (5.49) we establish that $\|r_1\| \leq C\|f'_1\| \leq C\|f'_0\|$. Taking into account also that $\|r_0\| \geq C\|f'_0\|$ we obtain $\|r_1\| \leq C\|r_0\|$. Consequently, $\|f''_{1c} - f''_{0c}\| \leq C\|r_0\|$. Using this we find that

$$|(e_1, r_0)| \leq C\|r_0\|^2\|r_1\|.$$

Suppose that estimates (5.9) and (5.52) hold with $0 \leq i, j \leq \tau < n-1$. Then since $(f'_{j+1}, r_j) = 0$, we have

$$|(f'_{\tau+1}, r_j)| = |(e_\tau + \dots + e_{j+1}, r_j)| \leq C\|r_j\|^2\|r_{j+1}\|, \quad 0 \leq j < \tau. \quad (5.54)$$

Hence using (5.53),

$$|(f'_{\tau+1}, r_j)| \leq C\|r_j\|\|f'_j\|\|f'_{j+1}\|. \quad (5.55)$$

If (5.54) is satisfied, then $\|f'_j\|\|f'_{j+1}\| \leq \lambda_\xi\|f'_{\tau+1}\|$. Making use of this inequality we obtain from (5.55) that

$$|(f'_{\tau+1}, r_j)| \leq \lambda_\xi\|r_j\|\|f'_{\tau+1}\|, \quad 0 \leq j < \tau.$$

Since $(f'_{\tau+1}, r_\tau) = 0$, we finally obtain

$$|(f'_{\tau+1}, r_j)| \leq \lambda_\xi\|r_j\|\|f'_{\tau+1}\|, \quad 0 \leq j < \tau. \quad (5.56)$$

Using estimates (5.27), (5.56) and (5.25) we obtain also:

$$\frac{(H_j f'_j, f'_{\tau+1})^2}{(H_j e_j, e_j)} \leq \frac{(f'_{\tau+1}, r_j)^2}{m\|r_j\|^2} \leq \lambda_\xi^2\|f'_{\tau+1}\|^2, \quad 0 \leq j \leq \tau, \quad (5.57)$$

$$\begin{aligned} \frac{|(H_j f'_{j+1}, f'_{\tau+1})(H_j f'_j, f'_{\tau+1})|}{(H_j e_j, e_j)} &\leq \frac{C\|f'_{j+1}\|\|f'_{\tau+1}\| |(f'_{\tau+1}, r_j)|}{m\|r_j\|^2} \\ &\leq \frac{\lambda_\xi\|f'_{\tau+1}\|^2\|f'_{j+1}\|}{\|r_j\|}. \end{aligned} \quad (5.58)$$

Taking into account that $\|r_j\| \geq C\|f'_j\|$, $0 \leq j \leq \tau$ and using (5.53), (5.49) we obtain

$$\|r_i\| \leq C\|f'_i\| \leq C\|f'_j\| \leq C\|r_j\|, \quad 0 \leq j \leq i \leq \tau. \quad (5.59)$$

Making use of (5.59) we have

$$\frac{\lambda_{\tau} \|f'_{\tau+1}\|^2 \|f'_{j+1}\|}{\|r_j\|} \leq \lambda_{\tau} \|f'_{\tau+1}\|^2, \quad 0 \leq j \leq \tau. \quad (5.60)$$

Using (5.57), (5.58) and (5.60) we establish in the same way as in the subsection on p. 111 that

$$(H_{\tau+1}f'_{\tau+1}, f'_{\tau+1}) \geq a_{\tau+1} \|f'_{\tau+1}\|^2.$$

Let us show now that

$$H_{\tau+1}e_j = r_j + \eta_{j, \tau+1}, \quad 0 \leq j \leq \tau \quad (5.61)$$

where

$$\|\eta_{j, \tau+1}\| \leq C \sum_{v=j+1}^{\tau} \frac{\|r_j\|^2 \|r_{j+1}\|}{\|r_v\|}, \quad 0 \leq j < \tau, \quad \|\eta_{\tau, \tau+1}\| = 0. \quad (5.62)$$

Indeed,

$$H_{s+1}e_j = H_s e_j + \frac{r_s(r_s, e_j)}{(r_s, e_s)} - \frac{(e_s, H_s e_j) H_s e_s}{(H_s e_s, e_s)}, \quad (5.63)$$

we have $H_{j+1}e_j = r_j$. Using estimates (5.52), which hold by assumption with $0 \leq s, j \leq \tau$, (5.25), (5.27), (5.47) and taking into account that $|\alpha_i| \leq C$, $0 \leq i \leq \tau$, we obtain with $s = j + 1$

$$\frac{\|(e_{j+1}, H_{j+1}e_j) H_{j+1}e_{j+1}\|}{(H_{j+1}e_{j+1}, e_{j+1})} = \frac{\|(e_{j+1}, r_j) H_{j+1}e_{j+1}\|}{(H_{j+1}e_{j+1}, e_{j+1})} \leq C \frac{\|r_j\|^2 \|r_{j+1}\|}{\|r_{j+1}\|}.$$

Noting also that $(r_{j+1}, e_j) = 0$ we obtain from (5.63):

$$H_{j+2}e_j = r_j + \eta_{j, j+2}, \quad \|\eta_{j, j+2}\| \leq C \frac{\|r_j\|^2 \|r_{j+1}\|}{\|r_{j+1}\|}.$$

Suppose that with a certain $j+1 < s \leq \tau$ we have

$$H_s e_j = r_j + \eta_{j, s}, \quad \|\eta_{j, s}\| \leq C \sum_{v=j+1}^{s-1} \frac{\|r_j\|^2 \|r_{j+1}\|}{\|r_v\|}.$$

Then because of the same conditions used with $s = j + 1$ we have:

$$\begin{aligned} \frac{\|(e_s, H_s e_j) H_s e_s\|}{(H_s e_s, e_s)} &\leq C \frac{\|r_j\|^2 \|r_{j+1}\|}{\|r_s\|} + C \sum_{v=j+1}^{s-1} \frac{\|r_j\|^2 \|r_{j+1}\|}{\|r_v\|}, \\ \frac{\|r_s(r_s, e_j)\|}{(r_s, e_s)} &\leq C \frac{\|r_j\|^2 \|r_{j+1}\|}{\|r_s\|}. \end{aligned}$$

Using these estimates in (5.63) we establish that

$$H_{s+1}e_j = r_j + \eta_{j, s+1}, \quad \|\eta_{j, s+1}\| \leq C \sum_{v=j+1}^s \frac{\|r_j\|^2 \|r_{j+1}\|}{\|r_v\|}.$$

Thus, by induction, we can consider (5.61) to hold.

We can prove now that estimates (5.52) hold with $i = \tau + 1$. Making use of (5.61) we have

$$\begin{aligned} (r_{\tau+1}, e_j) &= -\alpha_{\tau+1} (H_{\tau+1} f'_{\tau+1}, e_j) \\ &= -\alpha_{\tau+1} (f'_{\tau+1}, r_j + \eta_{j, \tau+1}). \end{aligned} \quad (5.64)$$

By (5.59) and (5.49), we have

$$\|r_v\| \geq C \|f'_\tau\| \geq C \|f'_{\tau+1}\|, \quad 0 \leq v \leq \tau. \quad (5.65)$$

Taking this into account and using (5.62), we obtain $|(f'_{\tau+1}, \eta_{j, \tau+1})| \leq C \|r_j\|^2 \|r_{j+1}\|$. Using this inequality and also estimates (5.54) in (5.64) and taking into account that $\alpha_{\tau+1} \leq C$, we find that

$$|(r_{\tau+1}, e_j)| \leq C \|r_j\|^2 \|r_{j+1}\|, \quad 0 \leq j \leq \tau. \quad (5.66)$$

Further,

$$\begin{aligned} (e_{\tau+1}, r_j) &= (r_{\tau+1}, e_j) + (r_{\tau+1}, (f''(x_{\tau+1} + \theta_{\tau+1} r_{\tau+1}) \\ &\quad - f''(x_j + \theta_j r_j)) r_j). \end{aligned} \quad (5.67)$$

By (5.65) and (5.53),

$$\|r_{\tau+1}\| \leq C \|r_j\|, \quad 0 \leq j \leq \tau.$$

Consequently, using (2.8) we have

$$\begin{aligned} \|f''(x_{\tau+1} + \theta_{\tau+1} r_{\tau+1}) - f''(x_j + \theta_j r_j)\| &\leq R (\|x_{\tau+1} - x_j\| \\ &\quad + \|r_{\tau+1}\| + \|r_j\|) \leq C \|r_j\|. \end{aligned}$$

Using this estimate and (5.66) in (5.67) we obtain:

$$|(e_{\tau+1}, r_j)| \leq C \|r_j\|^2 \|r_{j+1}\|, \quad 0 \leq j \leq \tau.$$

Thus one step of induction has been completed, i.e. it has been established that estimates (5.9) and (5.52) hold. The proof of the lemma is completed.

Theorem 5.2. *Let $f(x)$ be a twice continuously differentiable function and matrix $f''(x)$ satisfy conditions (5.3) and (2.8). If $f(x)$ is minimized by algorithm $\{(5.1), (4.48)\}$, then with any sufficiently large ξ the following estimate holds:*

$$\|x_{(\xi+1)n} - x_*\| \leq C \|x_{\xi n+1} - x_*\| \|x_{\xi n} - x_*\|. \quad (5.68)$$

Proof. By (5.45), estimate (5.68) is equivalent to

$$\|f'_n\| \leq C \|f'_1\| \|f'_0\|. \quad (5.69)$$

Suppose that estimate (5.69) with all sufficiently large ξ does not hold. Then there must exist an infinite subsequence $\{\xi_m\}$ such that for corresponding points the following inequalities hold:

$$\|f'_0\| \|f'_1\| \leq \lambda_{\xi_m} \|f'_n\|, \quad \lambda_{\xi_m} \rightarrow 0 \text{ as } \xi_m \rightarrow \infty. \quad (5.70)$$

Without loss of generality, it can be assumed that the subsequence $\{\xi_m\}$ coincides with the whole sequence $\xi = 0, 1, \dots$. Taking into account (5.49) we can ascertain that if (5.70) holds, estimates (5.51) hold too. Consequently, if we assume estimates (5.70) to be satisfied, then the requirements of theorem 5.2 provide for the fulfilment of the conditions of lemma 5.1. Thus, if (5.70) is fulfilled, the estimates (5.9) and (5.52) hold.

Taking this into account we have

$$\begin{aligned} |(f'_n, r_j)| &= |(e_{n-1} + \dots + e_{j+1}, r_j)| \\ &\leq C \|r_j\|^2 \|r_{j+1}\|, \quad 0 \leq j \leq n-2. \end{aligned}$$

Now in a way analogous to that used in establishing (5.56) we can show that

$$|(f'_n, r_j)| \leq \lambda_\xi \|f'_n\| \|r_j\|, \quad \lambda_\xi \rightarrow 0, \quad 0 \leq j \leq n-1. \quad (5.71)$$

Let us demonstrate now that if (5.51) holds, then the system $r_0, \dots, r_{n-1}$ is linearly independent. Note first of all that due to estimates (5.9), it follows from (5.51) that

$$\|r_i\| \|r_{i+1}\| \leq \lambda_\xi \|r_j\|, \quad 0 \leq i < j \leq n-1. \quad (5.72)$$

Making use of (5.72) and (5.47), estimates (5.52) can take the following form

$$\begin{aligned} |(r_i, e_j)| &\leq \lambda_\xi \|r_i\| \|r_j\| \leq \lambda_\xi \|r_i\| \|e_j\|, \quad \lambda_\xi \rightarrow 0, \\ &0 \leq i \neq j \leq n-1. \end{aligned} \quad (5.73)$$

We denote $\bar{r}_i = r_i / \|r_i\|$ and let

$$\|\varphi\| = \min_{\substack{n-1 \\ \sum_{i=0}^{n-1} |\beta_i| = 1}} \left\| \sum_{i=0}^{n-1} \beta_i \bar{r}_i \right\| = \left\| \sum_{i=0}^{n-1} \bar{\beta}_i \bar{r}_i \right\|.$$

Then

$$|(\varphi, e_j)| \geq \left| |\bar{\beta}_j (r_j, e_j)| - \left| \sum_{i \neq j} \bar{\beta}_i (\bar{r}_i, e_j) \right| \right|. \quad (5.74)$$

Since $\sum_{i=0}^{n-1} |\bar{\beta}_i| = 1$, we have $|\bar{\beta}_j| \geq \beta > 0$, at least for one index $j \in \overline{0, n-1}$. Then by (5.18) and (5.47) we have

$$|\bar{\beta}_j (r_j, e_j)| \geq C \|r_j\| \geq C \|e_j\|.$$

With $i \neq j$ making use of (5.73) we have

$$|\bar{\beta}_i (\bar{r}_i, e_j)| \leq \lambda_\xi |\bar{\beta}_i| \|\bar{r}_i\| \|e_j\| = \lambda_\xi |\bar{\beta}_i| \|e_j\|,$$

$$\lambda_\xi \rightarrow 0 \text{ as } \xi \rightarrow \infty.$$

Using the inequalities obtained in (5.74) we have with sufficiently large ξ $|(\varphi, e_j)| \geq C \|e_j\|$, i.e.

$$\|\varphi\| \geq C. \quad (5.75)$$

Hence, it follows that the system of vectors $r_0, \dots, r_{n-1}$ is linearly independent. Besides, using (5.75) it is easy to establish that the following statement holds: if $\psi_0, \dots, \psi_{n-1}$ is a system biorthogonal to $r_0, \dots, r_{n-1}$, then with sufficiently large ξ we have

$$\|r_i\| \|\psi_i\| \leq C, \quad 0 \leq i \leq n-1. \quad (5.76)$$

Finally, it can be ascertained that under conditions (5.71) and (5.76) the system of vectors $f'_n, r_0, \dots, r_{n-1}$ is also linearly independent. Indeed, suppose that

$$f'_n = \sum_{i=0}^{n-1} \gamma_i \psi_i = \sum_{i=0}^{n-1} (f'_n, r_i) \psi_i.$$

Then by (5.71) and (5.76) we obtain

$$\|f'_n\| \leq C \lambda_\xi \|f'_n\|.$$

Since $\lambda_\xi \rightarrow 0$, the last inequality cannot be satisfied with sufficiently great ξ . Hence, it follows that the system $f'_n, r_0, \dots, r_{n-1}$ is linearly independent.

Thus having assumed that estimate (5.68) does not hold with any $\xi \geq \xi_0$ (where ξ_0 is a certain sufficiently great number), we have proved that a system of $n+1$ vectors $f'_n, r_0, \dots, r_{n-1}$ in space E^n is linearly independent. However, this is impossible. Thus the initial assumption was wrong, i.e. estimate (5.68) holds.

The theorem is proved.

Discussion of Results

Thus we have made it clear that all of the methods studied in Sec. 4 can be applied for minimizing nonquadratic functions and that the convergence of the processes can be guaranteed for a class of functions that can be minimized by gradient methods. In the case where methods of conjugate directions are used for minimization of strongly convex functions, the rate of convergence proves not slower than superlinear.

The rate of convergence of methods of conjugate directions was established in a somewhat different manner than that of methods of other classes studied in the preceding sections: the sequence considered was $\{x_{\xi n}\}$ rather than $\{x_k\}$, i.e. actually we considered as one iteration a unified group of n usual iterations of the process $x_{\xi n}, x_{\xi n+1}, \dots, x_{\xi n+n-1}$. Speaking generally, the real rate of convergence of such processes may prove slower than of methods of dual

directions (Sec. 3) and the more so than of Newton's method (Sec. 2) (i.e. the decrease of the function value at each iteration $|f_{k+1} - f_k|$ in methods of the class under consideration may prove less than in methods of Secs. 2, 3 and the ratio $\|x_{k+1} - x_*\| / \|x_k - x_*\|$ greater). Thus, if for instance, in some algorithm we have

$$x_{k+1} - x_k = -D_k^{-1}f'_k \quad (5.77)$$

and in a method of conjugate directions

$$x_{(\xi+1)n} - x_{\xi n} = -D_{\xi n}^{-1}f'_{\xi n}$$

and $D_k = D_{\xi n} \rightarrow f''_k$, then this means that n iterations of the method of conjugate directions are equivalent, as to their convergence, to only one iteration of process (5.77). Nevertheless the rate of convergence of the methods of the class under consideration is practically rather fast and exceeds by far that of gradient methods.

At the same time, as mentioned in Sec. 4, the methods of conjugate directions differ but slightly from the gradient methods as to the labour per iteration.

The foregoing makes it possible to conclude that the methods of conjugate directions are of the most effective for solving minimization problems.

In this section we limited ourselves to the study of several concrete algorithms constructed in Sec. 4, though we could have studied also the properties of other algorithms that can be constructed according to the general scheme discussed in Sec. 4. However, the technique of studying other algorithms would not considerably differ from that used in Sec. 5. Indeed, the difference in the technique of proving theorem 5.1 amounts only to somewhat different ways of investigating the properties of matrix H_k . But in any method of the class under consideration, vectors u_k and v_k used for constructing H_{k+1} can be but various combinations of vectors r_k and $H_k^*e_k$ (see (4.32)), and the algorithms discussed in Secs. 4, 5 were chosen so as to use in constructing matrices H_k various combinations of these elements.

Using the results obtained, we now compare the properties of different algorithms in the minimization of nonquadratic functions.

The results of theorem 5.2 (estimate (5.68)) show that the rate of convergence of sequence $\{x_{\xi n}\}$ depends considerably on the properties of matrix $H_{\xi n}$. If, as $\xi \rightarrow \infty$,

$$H_{\xi n} \rightarrow (f''_{\xi n})^{-1}, \quad (5.78)$$

then

$$\frac{\|x_{\xi n+1} - x_*\|}{\|x_{\xi n} - x_*\|} \rightarrow 0$$

and the rate of convergence increases. This fact is practically of the greatest interest for algorithms having the property that in minimizing a quadratic function we have

$$H_n = A^{-1}. \quad (5.79)$$

Algorithms (4.48), (4.49), (4.52) belong to methods of this group. If in implementing one of such algorithms condition (5.78) is fulfilled, then, by the above considerations, it is expedient not to restore matrix H_k .

In method (4.70), property (5.79) is not fulfilled; therefore the variant of this method without restoration gives no advantage (we refer to the rate of convergence) over the variant with restoration. The same can be said also of other algorithms that have the property that in minimizing a quadratic function we have $H_n = H_0$ (for instance, methods (4.69), (4.71)) or H_n is close to H_0 (just this is the case of matrix H_n in method (4.70): its effect on the system of linearly independent vectors $e_0, \dots, e_{n-1}$ is the same as that of matrix H_0 , except on the vector e_{n-1}). Therefore it is not worth while to consider variants of such methods without restoration of matrix H_k .

However, the rate of convergence of methods (4.70), (4.71) will increase if we use, instead of the fixed matrix H_0 , a sequence of positive definite matrices $H_{\xi 0}$ which satisfy the condition

$$H_{\xi 0} \rightarrow (f''_{\xi n})^{-1}. \quad (5.80)$$

In cases where the requirements of lemma 5.1 are met, matrices $H_{\xi 0}$ which satisfy condition (5.80) can be constructed using vectors $r_{\xi n}, r_{\xi n+1}, \dots, r_{\xi n+n-1}$, by the formula

$$H_{(\xi+1)0} = \sum_{i=0}^{n-1} \frac{r_{\xi n+i} r_{\xi n+i}^*}{(r_{\xi n+i}, e_{\xi n+i})}.$$

In the light of the above considerations, the most effective methods of the class of methods of conjugate directions from the viewpoint of the rate of convergence in the minimization of strictly convex functions should be methods that have property (5.79).

In practice, deviations from this conclusion can of course occur in the sense that, for example, using method (4.70) the solution of the minimization problem (to a given accuracy) can be obtained after a smaller number of iterations, as compared to method (4.48), say. The fact, as we stressed many times, is that the rate of convergence of any method is affected by many additional factors, for instance, by errors in calculations, the choice of the values of α_k being made not precisely enough, and the sensitiveness of the methods to perturbations is different. Besides, the comparison of the rates of convergence makes sense only in a sufficiently small region about the

minimum, and at a point distant from it one can compare the effectiveness of different algorithms only on the ground of numerical experiments.

Many works published up to the present time (J.D. Pearson, J. Greenstadt, B.T. Polyak [2], H.G. Huang and A.V. Levy) contain results of numerical solution of various problems by methods of conjugate directions. The most comprehensive comparative analysis of the effectiveness of different algorithms is given in the last of the works named. On the whole, the results of numerical experiments confirm the conclusion that the most effective methods are those for which condition (5.79) is fulfilled. At the same time, method (4.71) proves more effective in the case where the matrix is restored after n iterations (as compared with a process without restoration). It seems that in practice method (4.70) should also be used with restoration of matrix H_k .

Finally, we dwell on problems that are involved in the choice of the step length in methods of the class under consideration. As was already mentioned, in methods of conjugate directions the step length is chosen under the condition that the minimum of the function is in the direction of motion. It was stressed many times that the main shortcoming of such a procedure is the necessity of performing a considerable amount of calculations of function values, this making the computational effort very considerable in problems in which the function evaluation requires much time. In some cases the selected method of choosing the step length is not practically suited if, for instance, the value of parameter α_k changes greatly at every step. This shortcoming of the method of conjugate directions was stressed in many works (C.G. Broyden [2], B.N. Pshenichny [3], W.C. Davidon [2], M.J.D. Powell [1], R. Fletcher [1], and others). To avoid the above shortcoming, these works consider methods in which the α_k value is chosen so that it guarantees only a certain degree of decrease of the function. However in other respects, the construction of these methods is based on the same ideas, which were described above (the work of B.N. Pshenichny [3] excepted).

The study of the properties of methods in which the choice of the step length is not connected with the finding of the function minimum along the direction of motion becomes much more difficult and the theoretical substantiation of many of them has not been performed even in the case of the minimization of a quadratic function.

From the viewpoint of the method of choosing the α_k value, methods of dual directions, Sec. 3, are preferable. The rate of convergence of these methods also proves faster. However, methods of dual directions require a greater storage capacity of the computer (as noted in Sec. 3, two $n \times n$ matrices must be stored); therefore using them, one can solve minimization problems but of smaller size. One can, though, use a smaller storage capacity of the computer by

choosing in methods of dual directions vectors r_k along the coordinate axes; however, in this case it is necessary to calculate the derivative twice at every iteration and this increases the amount of work required.

6. METHODS WITHOUT CALCULATING DERIVATIVES

Introductory Remarks

Until now we described minimization methods in which it was necessary at each iteration to calculate, besides the function $f(x)$, its gradient $f'(x)$ (methods of Secs. 1, 3, 4, 5), and in Newton's method (Sec. 2), moreover, the matrix of second derivatives $f''(x)$. Many times we stressed the fact that the calculation of second derivatives is often the most complicated and laborious part of the construction of the iterative process, and methods of Secs. 3-5 were worked out just with the aim of avoiding the calculation of second derivatives. However in many problems, the calculation of the gradient can also prove considerably more complicated than the evaluation of the function (in some cases it is impossible to obtain an analytical expression of $f'(x)$ at all). In such cases it is desirable to use methods which require only the function evaluation.

The calculation of a gradient by an analytical formula can be substituted by an approximate one, for instance, by using the finite differences approximation to partial derivatives. In this way one can construct modifications of the methods (discussed in the preceding sections) which involve only function evaluation. If we require a definite degree of accuracy of the approximation and impose certain additional requirements on the construction of an iterative process, then in most cases we can obtain that the properties of such modified methods (convergence, rate of convergence) approximate the properties of the original algorithms in which $f'(x)$, $f''(x)$ are evaluated by analytical expressions.

The study of methods without calculation of gradients is interesting also in another respect. In determining the accuracy of approximating the derivatives with which the properties of such algorithms coincide with those of the original methods, we find in fact the allowable calculation errors that do not lead to violations of the properties of algorithms (with the calculation of $f'(x)$, $f''(x)$).

In this section we are studying only those algorithms whose construction is based on methods of dual directions, Sec. 3; in this connection we retain for them the same name. Besides, we shall dwell on algorithms of another type in which the idea of the construction of conjugate directions is realized but without the calculation of the gradient or its approximation by finite differences.

Constructing Methods of Dual Directions

In these methods, successive approximations to the solution are constructed by the formula

$$x_{k+1} = x_k - \alpha_k D_k^{-1} g_k \quad (6.1)$$

where D_k is an $n \times n$ matrix, g_k is a vector. The scalar factor α_k which determines the step length, as distinct from the methods discussed before, can take positive as well as negative values; this depends on which direction $-D_k^{-1} g_k$ or $D_k^{-1} g_k$ is the direction of descent of function $f(x)$.

One can also use another approach and assume that $\alpha_k \geq 0$, but then the direction of motion should be either $p_k = -D_k^{-1} g_k$ or $p_k = D_k^{-1} g_k$ so that the following condition holds:

$$(f'_k, p_k) < 0 \quad (6.2)$$

We assume as we did in Sec. 3 that $f(x)$ is a twice continuously differentiable strongly convex function.

Constructing matrix D_k and vector g_k . Let us determine vectors θ_k and φ_k :

$$\begin{aligned} \theta_k &= \left(\frac{f(x_k + \mu_k v_1) - f(x_k)}{\mu_k}, \dots, \frac{f(x_k + \mu_k v_n) - f(x_k)}{\mu_k} \right), \\ \varphi_k &= \left(\frac{f(y_k + \mu_k v_1) - f(y_k)}{\mu_k}, \dots, \frac{f(y_k + \mu_k v_n) - f(y_k)}{\mu_k} \right), \\ \psi_k &= \varphi_k - \theta_k \end{aligned}$$

where $0 < |\mu_k| \leq \|r_k\|^t$, $t > 1$, y_k , r_k are elements of sequence (3.5), v_i is the unity vector of the corresponding axis.

Lemma 6.1. *Let $\{x_k\}$ be a bounded sequence, $\|x_{k+1} - x_k\| \rightarrow 0$ as $k \rightarrow \infty$ and matrix D_k with any $k \geq n - 1$ be defined by the following system of equations:*

$$D_k r_{k-i} = \psi_{k-i}, \quad i = 0, 1, \dots, n-1 \quad (6.3)$$

where r_{k-i} are elements of sequence (3.5). Then

$$\lim_{k \rightarrow \infty} \|D_k - f''_k\| = 0.$$

The proof of this lemma coincides in its essential features with that of lemma 3.1. We shall consider only the arising differences.

The components of vectors θ_k and φ_k can be written thus:

$$\begin{aligned} \theta_k^j &= \frac{\partial f}{\partial x^j} \Big|_{x=x_k + \theta_j \mu_k v_j}, \quad 0 \leq \theta_j \leq 1, \\ \varphi_k^j &= \frac{\partial f}{\partial x^j} \Big|_{x=y_k + \xi_j \mu_k v_j}, \quad 0 \leq \xi_j \leq 1, \\ j &= 1, \dots, n. \end{aligned}$$

Taking into account this and the continuity of second derivatives of the function, it is easy to ascertain that the following estimates hold:

$$\| \theta_k - f'(x_k) \| \leq C_1 |\mu_k| n^{1/2} \leq C_2 \|r_k\|^t, \quad (6.4)$$

$$\| \varphi_k - f'(y_k) \| \leq C_3 \|\mu_k\| n^{1/2} \leq C_4 \|r_k\|^t \quad (6.5)$$

where $C_2, C_4 < \infty$.

Let us write vector ψ_{k-i} in the form

$$\psi_{k-i} = f'(y_{k-i}) - f'(x_{k-i}) + (\varphi_{k-i} - f'(y_{k-i})) - (\theta_{k-i} - f'(x_{k-i})),$$

then denoting as before $e_{k-i} = f'(y_{k-i}) - f'(x_{k-i})$ we obtain

$$D_k r_{k-i} = e_{k-i} + (\varphi_{k-i} - f'(y_{k-i})) - (\theta_{k-i} - f'(x_{k-i})), \\ i = 0, 1, \dots, n-1. \quad (6.6)$$

Let us take $B_k = D_k - f''(x_k)$. Proceeding as in proving lemma 3.1 we obtain the following estimate:

$$\|B_k r_{k-i}\| \leq h_{k-i} \|r_{k-i}\| + \|\theta_{k-i} - f'(x_{k-i})\| + \|\varphi_{k-i} - f'(y_{k-i})\|$$

where $h_{k-i} \rightarrow 0$ as $k \rightarrow \infty$. Whence, due to (6.4) and (6.5), we have

$$\|B_k r_{k-i}\| \leq h_{k-i} \|r_{k-i}\| + C_5 \|r_{k-i}\|^t, \quad C_5 < \infty$$

or

$$\|B_k r_{k-i}\| \leq \tilde{h}_{k-i} \|r_{k-i}\|$$

where $\tilde{h}_{k-i} = h_{k-i} + C_5 \|r_{k-i}\|^{t-1} \rightarrow 0$ as $k \rightarrow \infty$.

The remaining part of the proof repeats the argument of lemma 3.1.

Let us now determine vector g_k :

$$g_k = \left(\frac{f(x_k + \rho_k \nu_1) - f(x_k)}{\rho_k}, \dots, \frac{f(x_k + \rho_k \nu_n) - f(x_k)}{\rho_k} \right) \quad (6.7)$$

where $|\rho_k| \leq |\mu_k|$ (if $\rho_k = \mu_k$, then $g_k = \theta_k$).

It is clear that the convergence and the rate of convergence of sequence (6.1) depend not only on the value of matrix D_k but also on how close vector g_k approximates gradient $f'(x_k)$. It will become clear from what follows that in order to guarantee a fast rate of convergence of sequence (6.1) to the solution, it is required that with any k the inequalities

$$0 < |\rho_k| \leq \xi_k \|p_k\| \quad (6.8)$$

where $\xi_k \rightarrow 0$ in an arbitrary manner as $k \rightarrow \infty$, be satisfied.

If at a certain iteration the chosen value of ρ_k does not satisfy conditions (6.8), it becomes necessary to take a smaller ρ_k , calculate a new vector g_k and then calculate a new vector p_k and check up

once more whether (6.8) is satisfied or not. Since $g_k \rightarrow f'_k$ as $|\rho_k| \rightarrow 0$ and at the same time $\|p_k\| \rightarrow \|D_k^{-1}f'_k\|$, and $\|D_k^{-1}f'_k\| > 0$ with any $x_k \neq x_*$ (matrix D_k^{-1} is nonsingular being the inverse of matrix D_k ; on the calculation of D_k^{-1} see the subsection on p. 111, calculation of vector p_k), then with sufficiently small values of ρ_k conditions (6.8) are satisfied.

Determining the direction of motion. This is made as follows. Setting a certain value of γ_0 (naturally this should be chosen sufficiently small), $f(x)$ is evaluated at points $x_k \pm \gamma_0 D_k^{-1}g_k$. If at one of these points the function value is less than $f(x_k)$, then the corresponding vector ($-D_k^{-1}g_k$ or $D_k^{-1}g_k$) is taken as p_k (condition (6.2) is satisfied since $f(x)$ is convex). However, if both function values are greater than $f(x_k)$, we reduce γ_0 until one of the function values becomes less than $f(x_k)$, and the corresponding vector is taken as p_k .

However, it can occur that with small values of γ the function does not decrease in either of the directions $\pm D_k^{-1}g_k$. This can mean that either we have not reached values of γ with which the function decreases or the condition $(f'_k, D_k^{-1}g_k) = 0$ is satisfied (it will be seen from what follows that such a case is possible only at the initial stage of the process and then, obviously, neither of the vectors $\pm D_k^{-1}g_k$ can be chosen, as p_k). In order to exclude such an occurrence it is necessary to calculate a new vector $g_{k,1}$ having changed ρ_k (but so that conditions (6.8) be satisfied), calculate a new vector $D_k^{-1}g_{k,1}$ and from a certain $\gamma \leq \gamma_0$ on, evaluate the function at points $x_k \pm \gamma D_k^{-1}g_{k,1}$ as well. If $x_k \neq x_*$, then one of the directions $\pm D_k^{-1}g_k$ or $\pm D_k^{-1}g_{k,1}$, is, of necessity, the direction of descent. The corresponding vector is then taken as p_k .

The algorithm of choosing the step. Let us choose α_k in the following way: suppose that

$$\bar{\alpha}_k = \min \left\{ 1, R \frac{|(g_k, p_k)|}{\|p_k\|^3} \right\} \quad (6.9)$$

where $0 < R < \infty$ and check the validity of the inequality

$$f(x) - f(x_k) \leq \varepsilon \alpha^2 \beta_k (g_k, p_k) \quad (6.10)$$

where $x = x_k + \alpha p_k$, $\beta_k = -\operatorname{sgn} (g_k, p_k)$, $0 < \varepsilon < \frac{1}{2}$.

If (6.10) holds with $\alpha = \bar{\alpha}_k$, then the value $\bar{\alpha}_k$ is taken as the required one, and if not, we reduce $\bar{\alpha}_k$ until (6.10) is satisfied; the value of α_k thus obtained is taken to be the one sought.

This method of choosing α_k presupposes, of course, that $(g_k, p_k) \neq 0$. If at a certain iteration we find that $(g_k, p_k) = 0$ (this can occur only at the initial stage of the process), then it is necessary to reduce ρ_k and calculate vector g_k anew.

We now study the properties of sequence (6.1) in constructing matrix D_k , vector g_k and parameter α_k by the method described above.

Theorem 6.1. *If $f(x)$ is a twice continuously differentiable function that satisfies conditions (2.4), matrix D_k with any $k \geq n - 1$ is defined by system (6.3), vector g_k is determined by expression (6.7) where p_k satisfies conditions (6.8) and α_k determined by the method described above, then for sequence (6.1) statements analogous to those proved in theorem 3.1 hold.*

Proof. In order to take advantage of lemma 6.1, it is necessary first of all to show that under the conditions of the theorem, condition $\|x_{k+1} - x_k\| \rightarrow 0$ holds for sequence (6.1).

Expanding function $f(x)$ into Taylor's series to the second-order terms in the region about point x_k we obtain:

$$f_{k+1} - f_k = \alpha_k \beta_k(g_k, p_k) \left[\frac{(f'_k, p_k)}{\beta_k(g_k, p_k)} + \frac{\alpha_k}{2} \frac{(f''_{kc} p_k, p_k)}{\beta_k(g_k, p_k)} \right]$$

where $x_{kc} = x_k + \theta(x_{k+1} - x_k)$, $0 \leq \theta \leq 1$. Since $\beta_k(g_k, p_k) < 0$, inequality (6.10) is satisfied if

$$\frac{(f'_k, p_k)}{\beta_k(g_k, p_k)} + \frac{\alpha_k}{2} \frac{(f''_{kc} p_k, p_k)}{\beta_k(g_k, p_k)} \geq \alpha_k \varepsilon$$

or which is the same thing

$$\frac{1}{\alpha_k} \frac{(f'_k, p_k)}{\beta_k(g_k, p_k)} + \frac{1}{2} \frac{(f''_{kc} p_k, p_k)}{\beta_k(g_k, p_k)} \geq \varepsilon. \quad (6.11)$$

Due to (6.2) and the choice of β_k

$$\frac{(f'_k, p_k)}{\beta_k(g_k, p_k)} > 0.$$

Consequently, with a certain $\alpha_k > 0$ the inequality (6.11) is satisfied and, therefore, (6.10) as well. This proves the possibility of choosing α_k by the method described above.

Thus, by (6.10), $f_{k+1} < f_k$. This means that $x_k \in S = \{x: f(x) \leq f(x_0)\}$ with any k and since $f(x)$ has a lower bound $f_k - f_{k+1} \rightarrow 0$. Hence, it follows from (6.10) that as $k \rightarrow \infty$

$$\alpha_k^2 |(g_k, p_k)| \rightarrow 0. \quad (6.12)$$

Since $\alpha_k \leq \bar{\alpha}_k$, it follows from (6.9) that

$$|(g_k, p_k)| \geq \frac{\alpha_k}{R} \|p_k\|^3.$$

With account taken of the last inequality, condition (6.12) implies that $\|x_{k+1} - x_k\| = \alpha_k \|p_k\| \rightarrow 0$ as $k \rightarrow \infty$. Consequently, the

conditions of the theorem provide for the satisfaction of the requirements of lemma 6.1 and therefore

$$\|D_k - f_k''\| \rightarrow 0. \quad (6.13)$$

We show now that under the conditions of the theorem as $k \rightarrow \infty$

$$\frac{(f_k', p_k)}{\beta_k(g_k, p_k)} \rightarrow 1. \quad (6.14)$$

We have

$$\frac{(f_k', p_k)}{\beta_k(g_k, p_k)} = \frac{1}{\beta_k} + \frac{(f_k' - g_k, p_k)}{\beta_k(g_k, p_k)} \leq \frac{1}{\beta_k} + \frac{\|f_k' - g_k\| \|p_k\|}{|(g_k, p_k)|}. \quad (6.15)$$

For vector g_k the following estimate (analogous to (6.4)) holds:

$$\|g_k - f_k'\| \leq C_6 |\rho_k| n^{1/2} = C_7 |\rho_k|. \quad (6.16)$$

Since

$$|(g_k, p_k)| = |(D_k p_k, p_k)| \quad (6.17)$$

it follows because of conditions (6.13) and (2.4), that from a certain iteration on,

$$|(g_k, p_k)| \geq m_1 \|p_k\|^2 \quad (6.18)$$

where $0 < m_1 \leq m$. Using estimates (6.8), (6.16) and (6.18), we find that from a certain iteration on, the following conditions will be fulfilled:

$$\frac{\|f_k' - g_k\| \|p_k\|}{|(g_k, p_k)|} \leq \frac{C_7 |\rho_k| \|p_k\|}{m_1 \|p_k\|^2} \leq \frac{C_7}{m_1} \xi_k \rightarrow 0.$$

Hence, it follows from (6.15) that from a certain iteration on, we have $\beta_k = +1$ (since the left-hand side of (6.15) is positive) and therefore condition (6.14) is really satisfied.

On set S the gradient $f'(x)$ is bounded: $\|f_k'\| \leq L$. Taking into account also that $|\rho_k| \leq \rho < \infty$, we find using (6.16) that $\|g_k\| \leq L_1$ with any k . By analogy with theorem 3.1 we can establish that with any $k \geq n - 1$ we have $\|D_k^{-1}\| \leq M_2$. Consequently,

$$\|p_k\| \leq \|D_k^{-1}\| \|g_k\| \leq C_8.$$

Using this estimate and inequality (6.18) we establish that with sufficiently great k

$$\frac{|(g_k, p_k)|}{\|p_k\|^3} \geq \frac{m_1 \|p_k\|^2}{\|p_k\|^3} \geq \frac{m_1}{\|p_k\|} \geq C_9 > 0. \quad (6.19)$$

Hence, it follows from (6.9) that from a certain k on, we shall have

$$\bar{\alpha}_k \geq \bar{\alpha} > 0. \quad (6.20)$$

It can be also easily ascertained using conditions (6.14) and (6.18) that with sufficiently great k inequality (6.11) and, therefore, (6.10) too are satisfied with values $\alpha \geq \bar{\alpha} > 0$. These inequalities together with estimate (6.20) show that from a certain k on, we find that

$$\alpha_k \geq C_{10} > 0.$$

Because of this estimate it follows from the condition $\alpha_k \|p_k\| \rightarrow 0$, the fulfilment of which was discussed above, that as $k \rightarrow \infty$

$$\|p_k\| \rightarrow 0. \quad (6.21)$$

Since $\|g_k\| = \|D_k p_k\| \leq M_1 \|p_k\|$, provided (6.21) is satisfied, we have

$$\|g_k\| \rightarrow 0. \quad (6.22)$$

In accordance with conditions (6.8), (6.16), (6.21) and (6.22), we can assert that as $k \rightarrow \infty$

$$\|f'(x_k)\| \rightarrow 0.$$

But this means, due to inequality (1.12) which holds for strongly convex functions, that sequence (6.1) converges to the solution.

Let us obtain an estimate of the rate of convergence of the method.

Due to conditions (6.17), (6.21) and the uniform convergence of second derivatives of the function on set S as $k \rightarrow \infty$, we have

$$\frac{(f''_{kc} p_k, p_k)}{|(g_k, p_k)|} \rightarrow 1.$$

Using this condition and (6.14) it is easy to ascertain that inequality (6.11) and, therefore, (6.10) as well with sufficiently great k are satisfied with $\alpha = 1$. From relations (6.19), with (6.21) fulfilled, it follows that

$$\frac{|(g_k, p_k)|}{\|p_k\|^3} \rightarrow \infty.$$

Therefore in choosing $\bar{\alpha}_k$ according to condition (6.9) from a certain k on, we have $\bar{\alpha}_k = 1$.

The above remarks show that from a certain iteration on, $\alpha_k \equiv 1$ and

$$x_{k+1} - x_k = -D_k^{-1} g_k.$$

At the same time there is a matrix $\tilde{D}_k^{-1}$ such that

$$x_{k+1} - x_k = -\tilde{D}_k^{-1} f'_k.$$

The sequence of matrices $\tilde{D}_k$ under conditions (6.8) and (6.16) can be chosen so that

$$\tilde{D}_k \rightarrow D_k. \quad (6.23)$$

In order to obtain (6.23) we can, for instance, assume that

$$\tilde{D}_k = D_k - \frac{(f'_k - g_k)}{\|x_{k+1} - x_k\|^2} (x_{k+1} - x_k)^*.$$

It is now easy to prove that sequence (6.1) converges at a super-linear rate. We proceed as in theorem 2.1 and establish that the following inequality holds:

$$\|x_{k+1} - x_*\| \leq \|\tilde{D}_k^{-1}\| \|\tilde{D}_k - f''_{kc}\| \|x_k - x_*\|.$$

Further using conditions (6.13), (6.23) and the continuity of second derivatives we ascertain that as $k \rightarrow \infty$ $\|\tilde{D}_k - f''_{kc}\| \rightarrow 0$ and quantity $\|\tilde{D}_k^{-1}\|$ has a bound. Hence, as $k \rightarrow \infty$ we have

$$\|x_{k+1} - x_*\| \leq \lambda_k \|x_k - x_*\| \quad (6.24)$$

where $\lambda_k \rightarrow 0$ and this proves that the rate of convergence of $\{x_k\}$ is superlinear.

The theorem is proved.

Remarks on the Implementation of Methods of Dual Directions

Various algorithms. The requirements which should be met by vectors r_k used in constructing matrix D_k are the same as those considered in constructing sequence (3.5). Therefore, all that was said in the subsection on p. 74 about the construction of various algorithms of type (3.4) holds for process (6.1).

Calculation of vector p_k . The results of the subsection on p. 76 are fully applicable here. Thus, basis $\bar{s}_{k+1}, \bar{s}_k, \dots, \bar{s}_{k-n+2}$, the dual of basis $\psi_{k+1}, \psi_k, \dots, \psi_{k-n+2}$, is constructed by the following formulas (analogous to (3.21)):

$$\bar{s}_{k+1} = \frac{s_{k-n+1}}{(s_{k-n+1}, e_{k+1})}, \quad \bar{s}_{k+1-j} = s_{k+1-j} - (s_{k+1-j}, e_{k+1}) \bar{s}_{k+1},$$

$$j = 1, \dots, n-1.$$

In this case in order to check that vectors $\psi_{k+1}, \psi_k, \dots, \psi_{k-n+2}$ are linearly independent, it suffices to calculate the scalar product (s_{k-n+1}, ψ_{k+1}) ; if $(s_{k-n+1}, \psi_{k+1}) \neq 0$, then vectors $\psi_{k+1}, \psi_k, \dots, \psi_{k-n+2}$ are linearly independent. But if we find that $(s_{k-n+1}, \psi_{k+1}) = 0$, then it is necessary to change either vector r_{k+1} or one of the vectors $\theta_{k+1}, \varphi_{k+1}$, thus changing vector ψ_{k+1} .

In practice, successive approximations should be constructed by the following formula (analogous to (3.25))

$$x_{k+1} = x_k - \alpha_k \sum_{i=0}^{n-1} (s_{k-i}, g_k) r_{k-i}. \quad (6.25)$$

The initial stage of the process. There are several ways of performing the first iterations of the process (with $k < n - 1$). For instance, the descent can be realized in one of the directions $\beta_k g_k$, $\beta_k = \pm 1$ choosing the sign of β_k so that $f(x)$ decreases.

In order to ensure uniformity of the iterative process (6.25), it can be started in a way analogous to that given in the subsection on p. 79.

Minimizing a quadratic form. Let $f(x) = \frac{1}{2} (Ax, x) + (b, x) + c$, where $(Ax, x) > 0$ with any $x \neq 0$. In this case it is easily ascertained that vector $\theta_k = g_k = f'(x_k)$, $\varphi_k = f'(y_k)$, $\psi_k = e_k$, i.e. $D_k = A_k$, and process (6.1) coincides with (3.4). Consequently (see the subsection on p. 79), process (6.1) allows to find the minimum of a quadratic function after n steps. It is necessary in this case to calculate $(n + 1)^2$ function values.

Choosing vector g_k . In method (6.1), besides approximating matrix $f''(x)$, we also substitute for gradient $f'(x)$ its finite differences analogue-vector g_k . In this case as was noted above in order to obtain a superlinear rate of convergence, conditions (6.8) are to be satisfied. If to satisfy these conditions we have to calculate at a certain iteration vector g_k several times, the amount of work required in the process increases (particularly, for a multidimensional space).

Note that if $\|p_k\| < \|p_{k-1}\|$ at each iteration, then one can choose $\|\rho_k\| = \|p_{k-1}\|^2$. It is very probable that with such a manner of choosing ρ_k , the right-hand one of the inequalities (6.8) will be satisfied, at least from a certain iteration on. Indeed, in the end we obtain bounds (6.24) on the rate of convergence.

The rate of convergence of a process estimated in this mode is usually slower than the quadratic one:

$$\|x_{k+1} - x_k\| \leq \|x_k - x_{k-1}\|^2, \quad \|x_k - x_{k-1}\| \rightarrow 0,$$

i.e. with bounds (6.24) usually $\|p_{k-1}\|^2 < \|p_k\|$ (recall that from a certain k on, we have $\alpha_k = 1$, i.e. $p_k = x_{k+1} - x_k$). Therefore, if sequence $\{\xi_k\}$ is chosen such that $\xi_k \rightarrow 0$ at a sufficiently slow rate, we can expect that with $\rho_k = \|p_{k-1}\|^2$ it will not be necessary to calculate g_k many times in order to satisfy (6.8). If, however, conditions (6.8) are not satisfied from the beginning (i.e. if we have to reduce ρ_k), this will suggest that the rate of convergence is close to the quadratic one.

In conclusion, note also that using the results of the subsection on p. 74 and of this section it is possible to establish the conditions of convergence of the modification of Newton's method that does not require calculation of derivatives.

Methods of Conjugate Directions

We consider a method of constructing conjugate directions which differs in the essence from the methods discussed in Sec. 4.

Let again

$$f(x) = \frac{1}{2} (Ax, x) + (b, x) + c$$

where $(Ax, x) > 0$ with any $x \neq 0$. Suppose that directions $p_1, \dots, p_m$, $m < n$ (not equal to zero) are A -orthogonal and $E^m(x_0)$ and $E^m(x_{0,m})$ are two different m -dimensional subspaces of space E^n that are formed by vectors $p_1, \dots, p_m$ and pass through points x_0 and $x_{0,m}$.

If x_m and $x_{m,m}$ are points of the minimum of $f(x)$ in subspaces $E^m(x_0)$ and $E^m(x_{0,m})$, then

$$\begin{aligned} (f'(x_m), p_i) &= 0, & (f'(x_{m,m}), p_i) &= 0, \\ i &= 1, 2, \dots, m. \end{aligned}$$

Consequently, $(f'(x_m) - f'(x_{m,m}), p_i) = 0$ or

$$(A(x_m - x_{m,m}), p_i) = 0, \quad i = 1, 2, \dots, m.$$

Thus if points of the minimum of $f(x)$ are determined in different subspaces formed by A -orthogonal directions $p_1, \dots, p_m$, then the direction $p_{m+1} = x_{m,m} - x_m$ proves to be conjugate to directions $p_1, \dots, p_m$.

The method described of constructing conjugate vectors does not require calculation of the gradient or its finite differences approximation. Let us now describe a concrete algorithm for the minimization of a quadratic function in which the construction of conjugate vectors is performed by the methods described.

We choose arbitrarily point x_0 and vector p_1 ; the m -th iteration of the algorithm ($m = 1, 2, \dots, n$) is performed as follows:

(1) Calculate point

$$x_m = x_{m-1} + \alpha_m p_m \tag{6.26}$$

where α_m is determined under the condition of the minimum function value:

$$f(\alpha) = f(x_{m-1} + \alpha p_m).$$

(2) Calculate point

$$x_{0,m} = x_m + r_m \quad (6.27)$$

where r_m is an arbitrary vector which is not a linear combination of vectors $p_1, \dots, p_m$ (below we shall dwell at some length on the question of the choice of r_m).

(3) Calculate points

$$x_{k,m} = x_{k-1,m} + \alpha_{k,m} p_k, \quad k = 1, \dots, m$$

where factor $\alpha_{k,m}$ is determined under the condition of the value of function $f(\alpha) = f(x_{k-1,m} + \alpha p_k)$ being minimum.

(4) Now calculate vector $p_{m+1} = x_{m,m} - x_m$. This is the end of the m -th iteration.

Vector r_m (in (6.27)) must not belong to the subspace $E^m(x_0)$ so that point $x_{0,m}$ would not belong to subspace $E^m(x_0)$. Since point x_m is the minimum point of $f(x)$ in subspace $E^m(x_0)$, it is clear that any vector $x - x_m$ in whose direction function $f(x)$ decreases does not belong to $E^m(x_0)$. Consequently, any direction of descent of $f(x)$ from point x_m can be taken as r_m . In particular, it is convenient to choose vector r_m along one of the coordinate axes; then if such a vector proves not to be the direction of descent, it is necessary to take as r_m a vector along another axis.

According to the results of Sec. 4, point x_n calculated by formula (6.26) is the minimum point of $f(x)$: $x_n = x_*$. In order to find point x_* we have to solve one-dimensional minimization problems (to determine factors α_m and $\alpha_{k,m}$) $1 + 2 + \dots + n = \frac{n+1}{2} n$ times.

Using this approach to the construction of the method of conjugate directions, one can construct various algorithms for the minimization of nonquadratic functions. Of course, in any algorithm of this kind, the directions $p_1, \dots, p_m$, $m \leq n$ will be no more conjugate (see the subsection on p. 103). However, we can expect that suitably worked out methods in a sufficiently small neighbourhood of the minimum point x_* (of a convex smooth function) will make possible the construction of vectors that are close enough as to their properties to the conjugate ones. Such algorithms may prove effective in minimizing nonquadratic functions.

We shall consider below an algorithm based on the above considerations.

Let $x_{1,0}$ be an arbitrary point and $v_{1,1}, \dots, v_{1,n}$ be an orthonormalized coordinate basis; the k -th iteration of the algorithm, $k = 1, 2, \dots$ consists of the following steps:

(1) For $i = 1, 2, \dots, n$ calculate

$$x_{k,i} = x_{k,i-1} + \alpha_{k,i} v_{k,i}$$

where $\alpha_{k,i}$ are determined under the condition that the function value is minimum:

$$f(\alpha) = f(x_{k,i-1} + \alpha v_{k,i}).$$

(2) Assume that

$$v_{k,n+1} = \frac{x_{k,n} - x_{k,0}}{\gamma_k}$$

where $\gamma_k = \|x_{k,n} - x_{k,0}\|$ and calculate point $x_{k,n+1} = x_{k,n} + \alpha_{k,n+1} v_{k,n+1}$, where $\alpha_{k,n+1}$ is determined under the condition that the value of the function

$$f(x_{k,n} + \alpha v_{k,n+1})$$

be minimum.

(3) Let $\alpha_{k,s} = \max\{\alpha_{k,i} : i = 1, 2, \dots, n\}$, Δ_k be a determinant whose columns are vectors $v_{k,1}, \dots, v_{k,n}$ and $\varepsilon > 0$ be an arbitrary small positive constant. If

$$\frac{\alpha_{k,s} \Delta_k}{\gamma_k} \geq \varepsilon,$$

we set that $v_{k+1,i} = v_{k,i}$ with $i \neq s$ and $v_{k+1,s} = v_{k,n+1}$; then we have

$$\Delta_{k+1} = \frac{\alpha_{k,s} \Delta_k}{\gamma_k}. \quad (6.28)$$

If we find that $\frac{\alpha_{k,s} \Delta_k}{\gamma_k} < \varepsilon$, we take $v_{k+1,i} = v_{k,i}$ for all $i = 1, 2, \dots, n$; then $\Delta_{k+1} = \Delta_k$.

(4) Take $x_{k+1,0} = x_{k,n+1}$; this is the end of the k -th iteration.

Equality (6.28) must be proved. However, preliminarily we shall discuss the algorithm proposed.

Let us consider a simplified variant of the algorithm whose k -th iteration is performed as follows:

(1) Construct points $x_{k,i}$, $i = 1, 2, \dots, n$ in the same way as in step (1) of the original algorithm.

(2) Calculate $x_{k,n+1} = x_{k,n} + \alpha_{k,n+1} v_{k,n+1}$, where $v_{k,n+1} = x_{k,n} - x_{k,0}$ and $\alpha_{k,n+1}$ provides the minimum of function $f(x_{k,n} + \alpha v_{k,n+1})$.

(3) Set $v_{k+1,i} = v_{k,i+1}$, $i = 1, 2, \dots, n$.

(4) Set $x_{k+1,0} = x_{k,n+1}$.

Let $k = 2$. Then we have

$$\begin{aligned} x_{2,0} &= x_{1,n+1} = x_{1,n} + \alpha_{1,n+1} v_{1,n+1}, \\ x_{2,n} &= x_{2,n-1} + \alpha_{2,n} v_{2,n} = x_{2,n-1} + \alpha_{2,n} v_{1,n+1}, \end{aligned}$$

i.e. points $x_{2,0}$ and $x_{2,n}$ are minimum points of $f(x)$ in the one-dimensional subspace (formed by vector $v_{1,n+1}$) which passes through two different points $x_{1,n}$ and $x_{2,n-1}$. If $f(x)$ is a quadratic function, then according to the foregoing the direction $v_{2,n+1} = x_{2,n} - x_{2,0}$ is found to be conjugate to the direction $v_{1,n+1} = v_{2,n}$. By similar reasoning, it can be ascertained that if with any $k = 1, 2, \dots, n$ vectors $v_{k,1}, \dots, v_{k,n}$ are linearly independent, then after the k -th iteration vectors $v_{k,n+1}, v_{k,n}, \dots, v_{k,n-k+2}$ will prove conjugate, i.e. after n iterations of the process we shall have constructed n conjugate vectors. However, it is impossible with this method of constructing vectors $v_{k,1}, \dots, v_{k,n}$ to guarantee their linear independence. Indeed, if with a certain k we have $\alpha_{k,1} = 0$ then, as is easily ascertained,

$$v_{k,n+1} = x_{k,n} - x_{k,0} = x_{k,n} - x_{k,1} = \sum_{i=2}^n \alpha_{k,i} v_{k,i},$$

i.e. at the $(k+1)$ -iteration the system of vectors $v_{k+1,i} = v_{k,i+1}$, $i = 1, 2, \dots, n$ is found to be linearly dependent. In this case it is not possible to construct a system of n conjugate vectors; this means that with the application of this simplified algorithm we cannot guarantee that a solution will be obtained even for a quadratic function. The more complicated steps (2) and (3) of the original algorithm are used just in order to avoid linear dependence of vectors $v_{k,i}$, $i = 1, 2, \dots, n$ (we find that $\Delta_k \geq \varepsilon$).

However, note that in minimizing a quadratic function with the aid of the original algorithm, it is impossible to guarantee that the problem will be solved after a finite number of iterations. Indeed, if we go over from the system of vectors $v_{k,1}, \dots, v_{k,n}$ to the system $v_{k+1,1}, \dots, v_{k+1,n}$, it can occur that one of the conjugate vectors already constructed is changed (see step (3)); therefore it is impossible to guarantee that n conjugate vectors will be obtained after a finite number of iterations. Besides, the system of vectors $v_{k,i}$ can remain unchanged in going over to the $(k+1)$ -iteration.

We show now that equality (6.28) holds:

$$\begin{aligned} \Delta_{k+1} &= \det [(v_{k+1,1}, \dots, v_{k+1,n})] \\ &= \det [(v_{k,1}, \dots, v_{k,s-1}, v_{k,n+1}, v_{k,s+1}, \dots, v_{k,n})]. \end{aligned}$$

But

$$v_{k,n+1} = \frac{1}{\gamma_k} (x_{k,n} - x_{k,0}) = \frac{1}{\gamma_k} \sum_{i=1}^n \alpha_{k,i} v_{k,i}.$$

Consequently

$$\Delta_{k+1} = \frac{\alpha_{k,s}}{\gamma_k} \det [(v_{k,1}, \dots, v_{k,s}, \dots, v_{k,n})] = \frac{\alpha_{k,s} \Delta_k}{\gamma_k}.$$

Thus with any k we find that $\Delta_k \geq \varepsilon$, and just this guarantees that vectors $v_{k,1}, \dots, v_{k,n}$ are linearly independent.

Let us study certain properties of this algorithm.

Theorem 6.2. *Let $f(x)$ be a continuously differentiable strictly convex function such that set $S = \{x: f(x) \leq f(x_{1,0})\}$ has a bound with an arbitrarily chosen point $x_{1,0}$. Then sequence*

$$\{x_{k,i}\}, \quad i = 0, 1, \dots, n, \quad k = 1, 2, \dots \quad (6.29)$$

constructed by the method described above converges to the minimum point of function $f(x)$.

Proof. The existence and uniqueness of minimum point x_* of function $f(x)$ under the conditions of the theorem follow from the results of lemmas 3.1 and 3.4 (Chap. I). Therefore it remains only to prove that sequence $\{x_{k,i}\}$ is convergent. Any point of sequence (6.29) $x_{k,i} \in S$ since $f(x_{k,i}) = \min_{\alpha} f(x_{k,i-1} + \alpha v_{k,i}) \leq f(x_{k,i-1})$ and $f(x_{k+1,0}) = f(x_{k,n+1}) \leq f(x_{k,n})$. Set S is bounded, i.e. (in E^n) it is compact. Consequently, on every infinite sequence of elements of this set it is possible to pick out a subsequence which converges to a certain element of S . If we consider sequence $\{x_{k,i}\}$ with a fixed $i = 0, 1, \dots, n$, then, by virtue of what has been stated, there is an infinite subsequence $\{x_{k_m,i}\}$ that converges to a point $x_i \in S$. At the same time since $f(x)$ has a lower bound, we have $f(x_{k_m,i+1}) \rightarrow f(x_{k_m,i})$. It follows, taking into account the continuity of $f(x)$, that the following equalities hold:

$$f(x_{i+1}) = \lim_{k_m \rightarrow \infty} f(x_{k_m,i+1}) = \lim_{k_m \rightarrow \infty} f(x_{k_m,i}) = f(x_i). \quad (6.30)$$

Let us demonstrate that $x_{i+1} = x_i$ for all $i = 0, 1, \dots, n-1$. By construction, $\|v_{k,i}\| = 1$ with any k and i , consequently, vectors $v_{k,i}$ can be considered to be elements of a unity sphere (of a bounded set) and therefore with any fixed $i = 1, \dots, n$ there is a subsequence $\{v_{k_m,i}\}$ that converges to a certain vector v_i . Since $x_{k,i+1} = x_{k,i} + \alpha_{k,i+1} v_{k,i+1}$ and $x_{k_m,i+1} \rightarrow x_{i+1}$, $x_{k_m,i} \rightarrow x_i$, $v_{k_m,i+1} \rightarrow v_{i+1}$ we have

$$x_{i+1} = x_i + \alpha_{i+1} v_{i+1}, \quad i = 0, 1, \dots, n-1$$

where $\alpha_{i+1} = \lim_{k_m \rightarrow \infty} \alpha_{k_m,i+1}$. Since condition $f(x_{k,i+1}) =$

$$= \min_{\alpha} f(x_{k,i} + \alpha v_{k,i+1}) \text{ is satisfied at point } x_{k,i+1}, \text{ we must have:}$$

$$f(x_{i+1}) = \min_{\alpha} f(x_i + \alpha v_{i+1}), \quad i = 0, 1, \dots, n-1, \quad (6.31)$$

i.e. the minimum of $f(x)$ in the direction v_{i+1} is attained at point x_{i+1} . But it follows from (6.30) that $f(x_{i+1}) = f(x_i)$. Since $f(x)$ is strictly convex, there is a unique minimum point in the direction v_{i+1} ; hence $x_{i+1} = x_i$.

Thus we find that $x_0 = x_1 = \dots = x_n$. Denoting this common point by $\tilde{x}$ we can rewrite condition (6.31) as follows

$$f(\tilde{x}) \leq f(\tilde{x} + \alpha v_i), \quad i = 1, 2, \dots, n \quad (6.32)$$

with any α . For a differentiable function these conditions are equivalent to the following ones:

$$(f'(\tilde{x}), v_i) = 0, \quad i = 1, 2, \dots, n. \quad (6.33)$$

Note now that since $\det[(v_{k,1}, \dots, v_{k,n})] \geq \varepsilon$, we have also $\det[(v_1, \dots, v_n)] \geq \varepsilon$. It follows that vectors $v_1, \dots, v_n$ are linearly independent. Taking this into account we have from (6.33) that $f'(\tilde{x}) = 0$. Due to the strict convexity of $f(\tilde{x})$, this means that $\tilde{x}$ is the minimum point of $f(x)$: $\tilde{x} = x_*$.

Thus we have proved that there is a subsequence $\{x_{k_m, i}\}$ which converges to point x_* . However, since with any fixed $i = 0, 1, \dots, n$ we have $f(x_{k+1, i}) \leq f(x_{k, i})$ and $f(x)$ has a lower bound, the following condition is satisfied:

$$\lim_{k \rightarrow \infty} f(x_{k, i}) = \lim_{k_m \rightarrow \infty} f(x_{k_m, i}) = f(x_i) = f(\tilde{x}).$$

It follows that with a fixed i the sequence $\{x_{k, i}\}$ is a minimizing one, consequently the sequence (6.29) is a minimizing one as well, and therefore, since there is only one minimum, the sequence converges to point x_* . The theorem is proved.

It is easy to ascertain that in proving that conditions (6.32) hold, we made no use of the fact that function $f(x)$ is differentiable, i.e. these inequalities hold also for a strictly convex continuous function. However, point $\tilde{x}$ —the limit point of sequence (6.29)—in this case can be not the minimum point of $f(x)$ (at the same time sequence (6.29) can have more than one limit point).

Discussion of Results

Note first of all that the field of application of the method of conjugate directions is broader than that of methods of dual directions; this is easily ascertained by comparing the requirements imposed on the function being minimized in theorems 6.1 and 6.2.

The properties of the method of conjugate directions under consideration have been as yet studied but insufficiently. Thus it is not yet clear what the rate of convergence of the algorithm is. Nevertheless, it is evidently slower (in minimizing functions of the same class) than that of methods of Sec. 5; this can be judged even by the fact that the algorithm under consideration does not guarantee the finding of the minimum of a quadratic form after n itera-

tions (and in fact, after a finite number of steps), i.e. it does not guarantee the construction of a system of n conjugate vectors after a finite number of iterations. Consequently, from the viewpoint of the rate of convergence, methods of dual directions with their superlinear rate of convergence have an advantage over the methods of conjugate directions.

Let us make an attempt to compare the amounts of computations at iterations of the algorithms studied.

In a method of type (6.1), it is necessary at each iteration to calculate the function value $n + 1$ or $2(n + 1)$ times for the construction of matrix D_k^{-1} and $n + 1$ times for the construction of vector g_k depending on the variant applied (see the subsection on p. 136); at the same time it can occur at some iterations that in determining g_k there is no need to perform new evaluations of the function (if $\rho_k = \mu_k$), or the amount of calculations can increase several times depending on how close the gradient is approximated by g_k . Besides, it is necessary to perform some more calculations of function values in order to choose the direction of motion and the step size.

In the method of conjugate directions it is necessary to calculate at each iteration the minimum of the function in the direction of motion $n + 1$ times. If we assume that in solving a one-dimensional minimization problem we have to calculate on the average 3 or 4 function values, then the amount of calculations at each iteration with the method being studied is about the same. It is not as yet clear though what accuracy the computation of the minimum in a direction of motion must be performed with in the method of conjugate directions so that the properties of the process be not violated. From the viewpoint of the influence on the convergence, the algorithm of α_k choice in process (6.1) is to be preferred.

On the whole, given the possibility of using methods of type (6.1), they must be more effective than the method of conjugate directions; however, it should be stressed once more that the field of application of the latter is broader.

Finally, it should be noted that in studying process (6.1) we made it practically clear that the calculation errors in determining vector e_k of the order of $O(\|r_k\|^t)$ (see (6.4), (6.5), (6.6)) and in determining vector $f'(x_k)$ of the order of $O(\xi_k \|p_k\|)$ (see (6.16)) do not violate the properties of process (3.4) (convergence, bounds on the rate of convergence). If we consider the variant of process (3.4) in which $r_{k+1} = x_{k+1} - x_k$, then we can obtain other expressions for estimating the errors. From a certain step on in process (3.4), $\alpha_k = 1$ and, consequently, we have

$$\|r_k\| = \|p_{k-1}\| = \|x_k - x_{k-1}\| = \|A_{k-1}^{-1} f'_{k-1}\| \geq m_1 \|f'_{k-1}\|.$$

Taking into account (1.12), we obtain $\|r_k\| \geq m_1 m \|x_{k-1} - x_*\|$.

Thus if $r_{k+1} = x_{k+1} - x_k$ the errors in the calculation of vectors e_k and f_k of the order of $O(\|x_{k-1} - x_*\|')$ and $O(\xi_k \|x_k - x_*\|)$ do not tell on the properties of process (3.4).

Bibliographic Notes

To Sec. 1. The idea of the gradient method was first stated by A. L. Cauchy. Methods of the gradient type were studied by L. V. Kantorovich [1], J. E. Kelley, B. T. Polyak [1], M. Altman, Yu. I. Lyubich, Yu. I. Lyubich and G. D. Maistrovsky. These works contain long lists of literature.

The variant of the method with the choice of the step size according to condition (1.2) which is described in this section is published for the first time.

The study of the rate of convergence of gradient methods given in this section is based on the results of the work of B. T. Polyak [1].

To Sec. 2. Newton's method for solving minimization problems and equations was studied by L. V. Kantorovich [2], L. V. Kantorovich and G. P. Aki-lov, and L. Collatz.

M. N. Yakovlev proved that the rate of convergence of the generalized method is superlinear if the step size is chosen from the condition that the function attains minimum in the direction of motion. A. A. Goldstein and J. F. Price, J. W. Daniel, Yu. M. Danilin [1, 2] studied Newton's method with adjustment of step length and used a mode of choice of the α_k value which did not involve the finding of the function minimum in the direction of motion.

To Sec. 3. This section is based on a paper by Yu. M. Danilin and B. N. Pshenichny [1].

To Sec. 4. The first of the methods of conjugate directions — the method of conjugate gradients—was proposed for solving problems of linear algebra by M. R. Hestenes and E. Stiefel. Another approach to the construction of methods of conjugate directions as applied to quadratic function minimization was proposed by W. C. Davidon [1] and developed by R. Fletcher, M.J.D. Powell and others.

In B. N. Pshenichny's algorithm [3] the construction of conjugate directions does not involve the finding of the function minimum in the direction of motion.

Many properties of conjugate directions are discussed by D. K. Faddeev and V. N. Faddeeva. The general method of constructing conjugate directions which is used in this section was worked out by H. Y. Huang. Some of the results are new, e.g. formula (4.45), method (4.63).

To Sec. 5. R. Fletcher and C. M. Reeves suggested the use of the method of conjugate gradients for the minimization of nonquadratic functions. The problems of the convergence and bounds on the rate of convergence of the method of conjugate gradients were studied by J. W. Daniel [1, 2], B. T. Polyak [2], G. D. Maistrovsky [1, 2], S. A. Smolyak. The convergence of method (4.48) and the bounds on the rate of convergence were established by M.J.D. Powell [3] (these results are described in the books by J. W. Daniel [1] and E. Polak [2]). The proof of the convergence of methods of conjugate directions described in this section is based on a paper by Yu. M. Danilin [4].

To Sec. 6. Methods of dual directions without calculating derivatives of the function are discussed in a paper by Yu. M. Danilin and B. N. Pshenichny [2]. Methods of conjugate directions were studied by C. S. Smith, M.J.D. Powell [2], W. I. Zangwill [1], J. W. Daniel [1]. The works named have been used in writing this section. A review of minimization methods without calculating derivatives was written by R. P. Brent.

CHAPTER III

METHODS OF CONSTRAINED FUNCTION MINIMIZATION

This chapter describes various methods of function minimization with constraints on the variables. The first section develops methods of solving problems of quadratic programming which is a subsidiary problem in many algorithms. The following sections describe the algorithms for solving problems of convex and nonconvex programming. Everywhere, if only feasible, the bounds on the rate of convergence are given.

1. PROBLEM OF QUADRATIC PROGRAMMING

Usually the problem of quadratic programming is understood to be the problem of the minimization of a quadratic function with linear constraints. Thus the problem of quadratic programming is the minimization of the function

$$f(x) = \frac{1}{2} (x, Cx) + (d, x) \quad (1.1)$$

with the following constraints

$$\begin{aligned} (a_i, x) - b_i &\leq 0, & i \in \mathcal{J}^-, \\ (a_i, x) - b_i &= 0, & i \in \mathcal{J}^0 \end{aligned} \quad (1.2)$$

where $x \in E^n$, $a_i \in E^n$, $i \in \mathcal{J}^- \cup \mathcal{J}^0$, $d \in E^n$, b_i are numbers, C is an $n \times n$ symmetric, positive definite matrix, i.e. $(x, Cx) \geq 0$ for all x , and $\mathcal{J}^-$ and $\mathcal{J}^0$ are finite sets of indices.

The basis of the numerical method of solving this problem is the method of conjugate gradients. The main idea of the application of this method to problem (1.1)-(1.2) is as follows.

Let x_0 be a point which satisfies constraints (1.2). We pick out among the constraints those which are satisfied as equalities. These constraints determine a certain face of the polyhedral set defined

by linear inequalities (1.2). We find the minimum of $f(x)$ on this face using the method of conjugate gradients. The point obtained is the solution of our problem or indicates a transition to a new face and then the procedure is repeated. Since the method of conjugate gradients minimizes function $f(x)$ after a finite number of steps, and the number of faces of the polyhedral set is limited, it is clear that an algorithm of this kind converges after a finite number of steps.

Operators of Projection

Let now $\mathcal{J} = \mathcal{J}^- \cup \mathcal{J}^0$ and $\mathcal{Y}$ be a subset of the set of indices $\mathcal{J}$. We form matrix $A_{\mathcal{Y}}$ whose rows are vectors a_i , $i \in \mathcal{Y}$ so that the matrix is $m \times n$ -dimensional, where m is the number of elements in set $\mathcal{Y}$.

Lemma 1.1. *If vectors a_i , $i \in \mathcal{Y}$ are linearly independent, then matrix $A_{\mathcal{Y}} A_{\mathcal{Y}}^*$ is nonsingular.*

Proof. Let $y \in E^m$ be a nonzero vector such that

$$A_{\mathcal{Y}} A_{\mathcal{Y}}^* y = 0. \quad (1.3)$$

Then

$$\begin{aligned} y^* A_{\mathcal{Y}} A_{\mathcal{Y}}^* y &= (A_{\mathcal{Y}}^* y)^* A_{\mathcal{Y}}^* y = (A_{\mathcal{Y}}^* y, A_{\mathcal{Y}}^* y) = \|A_{\mathcal{Y}}^* y\|^2 = 0, \\ A_{\mathcal{Y}}^* y &= 0 \end{aligned} \quad (1.4)$$

But $A_{\mathcal{Y}}^* y$ is just a linear combination of vectors a_i , $i \in \mathcal{Y}$ with coefficients y^i , $i = 1, \dots, m$, where y^i are components of vector y . By the assumption that a_i , $i \in \mathcal{Y}$ are linearly independent, this combination cannot be zero. Therefore, (1.4) and consequently (1.3) from which (1.4) was obtained are not true. Thus, matrix $A_{\mathcal{Y}} A_{\mathcal{Y}}^*$ can be made zero only by a zero vector, and this means that this matrix is nonsingular. Let us now define operator P :

$$P = A_{\mathcal{Y}}^* (A_{\mathcal{Y}} A_{\mathcal{Y}}^*)^{-1} A_{\mathcal{Y}}. \quad (1.5)$$

It is easily seen that operator P has the following properties:

$$PP = P, \quad (1.6)$$

$$P^* = P, \quad (1.7)$$

$$P(I - P) = (I - P)P = 0. \quad (1.8)$$

Operator P is the operator of orthogonal projection into a subspace spanned by vectors a_i , $i \in \mathcal{Y}$.

Indeed, for any vector $x \in E^n$

$$x = Px + (I - P)x.$$

Further by (1.7), (1.8),[‡]

$$(Px, (I - P)x) = (x, P^*(I - P)x) = 0$$

and so Px and $(I - P)x$ are components of the orthogonal resolution of vector x . Moreover

$$Px = A_{\mathcal{Y}}^* u = \sum_{i \in \mathcal{Y}} a_i u^i$$

where vector $u \in E^m$ with its components u^i is defined by formula $u = (A_{\mathcal{Y}} A_{\mathcal{Y}}^*)^{-1} A_{\mathcal{Y}} x$.

The expression for vector Px shows that it is wholly in the subspace spanned by vectors a_i , $i \in \mathcal{Y}$.

Note now that

$$A_{\mathcal{Y}}(I - P) = A_{\mathcal{Y}} - (A_{\mathcal{Y}} A_{\mathcal{Y}}^*) (A_{\mathcal{Y}} A_{\mathcal{Y}}^*)^{-1} A_{\mathcal{Y}} = 0. \quad (1.9)$$

Therefore, for any $x \in E^n$, vector $y = (I - P)x$ satisfies the system of equations $A_{\mathcal{Y}} y = 0$.

Minimization of a Quadratic Function in a Subspace

Suppose now that we have to minimize a quadratic function $f(x)$ defined by (1.1) with the constraints

$$(a_i, x) - b_i = 0, \quad i \in \mathcal{Y}. \quad (1.10)$$

We assume that vectors a_i , $i \in \mathcal{Y}$ are linearly independent.

Let x_0 be a point which satisfies (1.10).

Note that if we denote by $b_{\mathcal{Y}}$ a vector whose components are b_i , $i \in \mathcal{Y}$, then the system of equations (1.10) can be written in the form $A_{\mathcal{Y}} x - b_{\mathcal{Y}} = 0$ so that $A_{\mathcal{Y}} x_0 - b_{\mathcal{Y}} = 0$.

We now introduce a new variable y defined as follows:

$$x = x_0 + (I - P)y \quad (1.11)$$

and consider the quadratic function

$$\varphi(y) = f(x_0 + (I - P)y).$$

The gradients of functions $\varphi(y)$ and $f(x)$, according to the rules of differentiation of a composite function and the symmetry of operator P , are related as follows:

$$\varphi'(y) = (I - P)f'(x) \quad (1.12)$$

where x and y are connected by (1.11).

Lemma 1.2. *Let $\bar{y}$ be the point of absolute minimum of function $\varphi(y)$. Then the corresponding point*

$$\bar{x} = x_0 + (I - P)\bar{y}$$

is the minimum point of function $f(x)$ with constraints (1.10).

Proof. At point $\bar{y}$ the gradient of function $\varphi(y)$ becomes zero: $\varphi'(\bar{y}) = 0$. Therefore, by (1.12),

$$(I - P) f'(\bar{x}) = 0$$

or

$$f'(\bar{x}) - A_y^* (A_y A_y^*)^{-1} A_y f'(\bar{x}) = 0.$$

Taking $u = -(A_y A_y^*)^{-1} A_y f'(\bar{x})$, we obtain

$$f'(\bar{x}) + A_y^* u = 0. \quad (1.13)$$

Using (1.9) we obtain also

$$A_y \bar{x} = A_y x_0 + A_y (I - P) \bar{y} = A_y x_0 = b_y,$$

i.e. $\bar{x}$ satisfies conditions (1.10).

Thus $\bar{x}$ is the feasible point and at this point conditions (1.13) are satisfied, which are necessary and sufficient for $\bar{x}$ to be the minimum point of $f(x)$ with conditions (1.10). The lemma is proved.

Lemma 1.2 shows that the problem under consideration can be reduced to the minimization of quadratic function $\varphi(y)$ without constraints. To minimize $\varphi(y)$, we apply the method of conjugate gradients (Chap. II, Sec. 4):

$$y_0 = 0, \quad p_1 = -\varphi'(0),$$

$$y_{k+1} = y_k + \alpha_{k+1} p_{k+1},$$

$$p_{k+1} = -\varphi'(y_k) + \frac{\|\varphi'(y_k)\|^2}{\|\varphi'(y_{k-1})\|^2} p_k.$$

The quantity α_{k+1} in these formulas is calculated as follows

$$\alpha_{k+1} = -\frac{(\varphi'(y_k), p_{k+1})}{(p_{k+1}, (I - P) C (I - P) p_{k+1})}$$

since it is easy to ascertain that the matrix which determines the quadratic term of function $\varphi(y)$ has the form

$$(I - P) C (I - P).$$

These formulas determine the process involving the additional variables y . It is, however, expedient to go over to the original variables x . We preliminarily prove that the following relation holds:

$$(I - P) p_k = p_k. \quad (1.14)$$

Indeed, with $k = 1$ we have:

$$\begin{aligned} (I - P) p_1 &= -(I - P) \varphi'(0) = -(I - P) (I - P) f'(x_0) \\ &= -(I - P) f'(x_0) = -\varphi'(0) = p_1 \end{aligned}$$

where we have made use of (1.12) and of the fact that

$$(I - P)(I - P) = I - P - (I - P)P = I - P.$$

Now suppose that (1.14) holds for k and prove that it holds for $k + 1$, where we have again made use of (1.12) and (1.14):

$$\begin{aligned} (I - P)p_{k+1} &= -(I - P)\varphi'(y_k) + \frac{\|\varphi'(y_k)\|^2}{\|\varphi'(y_{k-1})\|^2}(I - P)p_k \\ &= -(I - P)(I - P)f'(x_k) + \frac{\|\varphi'(y_k)\|^2}{\|\varphi'(y_{k-1})\|^2}p_k \\ &= -(I - P)f'(x_k) + \frac{\|\varphi'(y_k)\|^2}{\|\varphi'(y_{k-1})\|^2}p_k = p_{k+1}. \end{aligned}$$

It follows now from (1.11) that

$$x_{k+1} = x_0 + (I - P)y_{k+1},$$

$$x_{k+1} = x_k + (I - P)(y_{k+1} - y_k) = x_k + (I - P)\alpha_{k+1}p_{k+1},$$

i.e.

$$x_{k+1} = x_k + \alpha_{k+1}p_{k+1}.$$

Let us transform the formula for p_{k+1} , using (1.12):

$$p_{k+1} = -(I - P)f'(x_k) + \frac{\|(I - P)f'(x_k)\|^2}{\|(I - P)f'(x_{k-1})\|^2}p_k.$$

The formula for α_{k+1} now takes the following form:

$$\alpha_{k+1} = -\frac{((I - P)f'(x_k), p_{k+1})}{((I - P)p_{k+1}, C(I - P)p_{k+1})} = -\frac{(f'(x_k), p_{k+1})}{(p_{k+1}, Cp_{k+1})}.$$

Theorem 1.1. *The problem of the minimization of quadratic function $f(x)$ with constraints (1.10), given the initial point x_0 which satisfies (1.10), is solved after a finite number of steps by the following process:*

$$\begin{aligned} p_1 &= -(I - P)f'(x_0), \\ x_{k+1} &= x_k + \alpha_{k+1}p_{k+1}, \\ p_{k+1} &= -(I - P)f'(x_k) + \frac{\|(I - P)f'(x_k)\|^2}{\|(I - P)f'(x_{k-1})\|^2}p_k, \\ \alpha_{k+1} &= -\frac{(f'(x_k), p_{k+1})}{(p_{k+1}, Cp_{k+1})}, \quad k = 0, \quad 1, \dots \end{aligned}$$

The proof of this theorem was given practically in the argument used in deriving the formulas of the process.

Remark. As we know (Chap. II, Sec. 4), if the method of conjugate gradients is applied to a quadratic function with a singular matrix C , then the process converges after a number of steps not exceeding $n - l$, where l is the number of zero eigenvalues of matrix C . In

minimizing $\varphi(y)$ we applied this method to a function whose matrix was $(I - P)C(I - P)$. But since $A_{\mathcal{Y}}(I - P) = 0$, that is, $(I - P)A_{\mathcal{Y}}^* = 0$, we have $(I - P)a_i = 0$, $i \in \mathcal{Y}$. Therefore, in the case under consideration the number of zero eigenvalues of matrix $(I - P)C(I - P)$ is not less than m , where m is the number of a_i , $i \in \mathcal{Y}$. Therefore, the process suggested either converges to the minimum point or shows no lower bound of quadratic function $f(x)$ with constraints (1.10) after a number of steps not exceeding $n - m$.

Algorithm of General Problem of Quadratic Programming

Let us now return to the general problem (1.1), (1.2). For each point x which satisfies (1.2) we set

$$\mathcal{Y}(x) = \{i: (a_i, x) - b_i = 0, \quad i \in \mathcal{J}^- \cup \mathcal{J}^0\}.$$

In what follows we assume that the following condition of non-degeneracy is fulfilled: with any x vectors a_i , $i \in \mathcal{Y}(x)$ are linearly independent.

We now propose the algorithm for solving the problem.

Let x_0 be an arbitrary point which satisfies (1.2) and is the first approximation. Take a set of indices $\mathcal{Y}_0 = \mathcal{Y}(x_0)$ and construct operator $P_{\mathcal{Y}_0}$:

$$P_{\mathcal{Y}_0} = A_{\mathcal{Y}_0}^* (A_{\mathcal{Y}_0} A_{\mathcal{Y}_0}^*)^{-1} A_{\mathcal{Y}_0}.$$

Calculate the quantities

$$u_0 = -(A_{\mathcal{Y}_0} A_{\mathcal{Y}_0}^*)^{-1} A_{\mathcal{Y}_0} f'(x_0), \quad (I - P_{\mathcal{Y}_0}) f'(x_0) = f'(x_0) + A_{\mathcal{Y}_0}^* u_0.$$

There are two possible cases:

(1) $(I - P_{\mathcal{Y}_0}) f'(x_0) = 0$. Here

$$f'(x_0) + A_{\mathcal{Y}_0}^* u_0 = 0 \tag{1.15}$$

and point x_0 is the minimum point of $f(x)$ on the face defined by the system of equations

$$(a_i, x) - b_i = 0, \quad i \in \mathcal{Y}_0$$

(see Chap. I, Sec. 3).

If there are no negative components among u_0^i , components of vector u_0 , $i \in \mathcal{Y}(x_0) \cap \mathcal{J}^-$, then (see Chap. I, Sec. 3) point x_0 is the solution of the primal problem (1.1), (1.2), for in this case (1.15) are the necessary and sufficient conditions for the minimum of function $f(x)$ with constraints (1.2).

Suppose now that there is an index $j \in \mathcal{Y}(x_0) \cap \mathcal{Y}^-$ such that $u_0^j < 0$. Construct a new set of indices $\mathcal{Y}'_0$ by deleting index j . We apply the method of conjugate gradients described in the subsection on p. 148 to solving the problem of minimization of $f(x)$ with constraints

$$(a_i, x) - b_i = 0, \quad i \in \mathcal{Y}'_0. \quad (1.16)$$

However, in applying the method of conjugate gradients, the process must not transgress the limits (1.2). Therefore at every step of the algorithm the following check should be made. Compute the quantity

$$\bar{\alpha}_{k+1} = \min_i \frac{b_i - (a_i, x_k)}{(a_i, p_{k+1})} \quad (1.17)$$

where the minimum is taken over all i for which $(a_i, p_{k+1}) > 0$. In this formula x_k is the point just constructed by the algorithm and p_{k+1} is the conjugate direction at this point.

Let now α_{k+1} be the corresponding step length in the method of conjugate gradients. If $\alpha_{k+1} < \bar{\alpha}_{k+1}$, then $x_{k+1} = x_k + \alpha_{k+1}p_{k+1}$ and the process goes on. If however $\alpha_{k+1} \geq \bar{\alpha}_{k+1}$, then $x_{k+1} = x_k + \bar{\alpha}_{k+1}p_{k+1}$ and the process stops.

Thus, either we find the minimum point of $f(x)$ under conditions (1.16) or the process will be truncated when $\alpha_{k+1} \geq \bar{\alpha}_{k+1}$. In both cases we take the point obtained to be the initial point and proceed using the new point as we did with the initial one, x_0 .

(2) $(I - P_{\mathcal{Y}_0}) f'(x_0) \neq 0$.

In this case we apply the method of conjugate gradients to solving the problem of minimization of $f(x)$ with constraints

$$(a_i, x) - b_i = 0, \quad i \in \mathcal{Y}_0 \quad (1.18)$$

starting at point x_0 . At every step, as before, a check is made whether the points obtained are feasible or not, i.e. we calculate $\bar{\alpha}_{k+1}$ by formulas (1.17) and apply the process of conjugate gradients until either we find the minimum point of $f(x)$ with constraints (1.18) or the condition $\alpha_{k+1} \geq \bar{\alpha}_{k+1}$ is satisfied and the point $x_{k+1} = x_k + \bar{\alpha}_{k+1}p_{k+1}$ obtained. In both cases we take the point obtained as the initial one and repeat at it the operations performed with x_0 .

Let us substantiate the convergence of the method after a finite number of steps. We must first of all show that in case (1) as well as in case (2) a successful step will be made, i.e. we move from point x_0 to a new point at which the value of function $f(x)$ will be strictly less than $f(x_0)$.

New points are obtained by the method of conjugate gradients and in this method the function decreases at each step. Therefore, the only thing we have to show is that $\bar{\alpha}_{k+1} > 0$ always, i.e. con-

straints (1.2) permit to take a nonzero step in the direction chosen, p_{k+1} , and besides that in case (1) point x_0 is not the minimum point of $f(x)$ with constraints (1.16), for if it were so the method of conjugate gradients would not have moved the process from point x_0 .

Let us prove several subsidiary lemmas.

Lemma 1.3. *Vector $p_1 = -(I - P_{\mathcal{Y}_0}) f'(x_0)$ is the solution of the problem of minimizing function*

$$\varphi(p) = (f'(x_0), p) + \frac{1}{2} \|p\|^2$$

with constraints

$$A_{\mathcal{Y}_0} p = 0. \quad (1.19)$$

Proof. Indeed, by (1.9), p_1 satisfies (1.19). Moreover, $\varphi'(p) = p + f'(x_0)$. Therefore,

$$\begin{aligned} \varphi'(p_1) &= p_1 + f'(x_0) = -(I - P_{\mathcal{Y}_0}) f'(x_0) + f'(x_0) \\ &= P_{\mathcal{Y}_0} f'(x_0) = -A_{\mathcal{Y}_0}^* u_0. \end{aligned}$$

Hence,

$$\varphi'(p_1) - A_{\mathcal{Y}_0}^* u_0 = 0. \quad (1.20)$$

The last expression is the necessary and sufficient condition for convex function $\varphi(p)$ to attain its minimum at point p_1 with constraints (1.19). The lemma is proved.

We formulate now a problem which is the dual of the problem of minimizing $\varphi(p)$ with constraints (1.19). According to the rules stated in Sec. 3 of Chap. I, we have to find the minimum of function $\varphi(p) + u^* A_{\mathcal{Y}_0} p$. Differentiating with respect to p and equating the derivatives to zero, we obtain $p + f'(x_0) + A_{\mathcal{Y}_0}^* u = 0$, i.e.

$$p = -f'(x_0) - A_{\mathcal{Y}_0}^* u.$$

Substituting this expression for p we obtain that

$$\min_p \{\varphi(p) + u^* A_{\mathcal{Y}_0} p\} = -\frac{1}{2} \|f'(x_0) + A_{\mathcal{Y}_0}^* u\|^2.$$

Thus the dual problem consists in finding over all possible vectors u the minimum of the function

$$\varphi^*(u) = -\frac{1}{2} \|f'(x_0) + A_{\mathcal{Y}_0}^* u\|^2.$$

Now differentiating $\varphi^*(u)$ and equating the derivatives to zero, we can easily ascertain that vector

$$u_0 = -(A_{\mathcal{Y}_0} A_{\mathcal{Y}_0}^*)^{-1} A_{\mathcal{Y}_0} f'(x_0)$$

is the solution of the dual problem, i.e. maximizes $\varphi^*(u)$. Recall that u_0^i , $i \in \mathcal{Y}_0$ are the components of u_0 . Thus, vector u_0 is the vector of Lagrange's multipliers in the problem of the minimization of $\varphi(p)$ with constraints (1.19). Besides, we obtain that the value of the minimum of $\varphi(p)$ with constraints (1.19) and of the maximum of $\varphi^*(u)$ over u , which is the same by the duality theorems, is equal to

$$-\frac{1}{2} \|f'(x_0) + A_{\mathcal{Y}_0}^* u_0\|^2 \text{ or } -\frac{1}{2} \|(I - P_{\mathcal{Y}_0}) f'(x_0)\|^2.$$

Lemma 1.4. *Let matrix $A_{\mathcal{Y}'_0}$ be formed from $A_{\mathcal{Y}_0}$ by deleting the row with index j for which $u_0^j < 0$ and let $(I - P_{\mathcal{Y}_0}) f'(x_0) = 0$. Then, vector $\bar{p}_1 = -(I - P_{\mathcal{Y}'_0}) f'(x_0)$ is not zero and $(a_j, \bar{p}_1) < 0$.*

Proof. Vector $\bar{p}_1$ can be written in the following form:

$$\bar{p}_1 = -(f'(x_0) + A_{\mathcal{Y}'_0}^* v), \quad v = -(A_{\mathcal{Y}'_0} A_{\mathcal{Y}'_0}^*)^{-1} A_{\mathcal{Y}'_0} f'(x_0).$$

If $\bar{p}_1 = 0$, then $f'(x_0) + A_{\mathcal{Y}'_0}^* v = 0$. But on the other hand, by assumptions,

$$(I - P_{\mathcal{Y}_0}) f'(x_0) = f'(x_0) + A_{\mathcal{Y}_0}^* u_0 = 0. \quad (1.21)$$

Subtracting from (1.21) the first equality, we obtain

$$A_{\mathcal{Y}_0}^* u_0 - A_{\mathcal{Y}'_0}^* v = u_0^j a_j + \sum_{i \neq j} (u_0^i - v^i) a_i = 0$$

which, as $u_0^j \neq 0$, is impossible since vectors a_i , $i \in \mathcal{Y}_0$ are linearly independent. We now prove the second part of the lemma.

Rewrite (1.21) in the component form:

$$f'(x_0) + \sum_{i \neq j} u_0^i a_i + (-u_0^j)(-a_j) = 0. \quad (1.22)$$

Note that $-u_0^j > 0$ since $u_0^j < 0$. Consider the problem of minimizing $\varphi(p) = (p, f'(x_0)) + \|p\|^2/2$ with constraints

$$(a_i, p) = 0, \quad i \in \mathcal{Y}'_0, \quad -(a_j, p) \leq 0. \quad (1.23)$$

Since $\varphi'(p) = f'(x_0) + p$, we have $\varphi'(0) = f'(x_0)$, and therefore, (1.22) is the necessary and sufficient condition for the point $p = 0$ to be the solution of the problem of minimization of $\varphi(p)$ with constraints (1.23). On the other hand, by lemma 1.3, $\bar{p}_1$ is the solution of the problem of minimization of $\varphi(p)$ with constraints $A_{\mathcal{Y}'_0}^* p = 0$ or in the component form

$$(a_i, p) = 0, \quad i \in \mathcal{Y}'_0. \quad (1.24)$$

Suppose that $(a_j, \bar{p}_1) \geq 0$. Since $\bar{p}_1$ satisfies constraints (1.24), it satisfies all constraints (1.23) too. But

$$\begin{aligned} (f'(x_0), \bar{p}_1) &= -(f'(x_0), (I - P_{\mathcal{Y}'_0}) f'(x_0)) \\ &= -(P_{\mathcal{Y}'_0} f'(x_0) + (I - P_{\mathcal{Y}'_0}) f'(x_0), (I - P_{\mathcal{Y}'_0}) f'(x_0)) \\ &= -((I - P_{\mathcal{Y}'_0}) f'(x_0), (I - P_{\mathcal{Y}'_0}) f'(x_0)) \\ &= -\|\bar{p}_1\|^2. \end{aligned}$$

Therefore,

$$\varphi(\bar{p}_1) = (f'(x_0), \bar{p}_1) + \frac{1}{2} \|\bar{p}_1\|^2 = -\frac{1}{2} \|\bar{p}_1\|^2 < 0.$$

The last inequality contradicts the fact that the minimum value of $\varphi(p)$ with constraints (1.23) is attained with $p = 0$ and is equal to zero. This contradiction shows that $(a_j, \bar{p}_1) < 0$. The lemma is proved.

We return now to the algorithm constructed. Let us consider case (1) and let point x_0 be not the solution of the problem of quadratic programming. According to the algorithm, we should apply the method of conjugate gradients in order to minimize function $f(x)$ with constraints (1.16). In accordance with the formulas of the method, the first step is made in the direction of vector

$$\bar{p}_1 = -(I - P_{\mathcal{Y}'_0}) f'(x_0).$$

By lemma 1.4, $\bar{p}_1 \neq 0$ and consequently point x_0 is not the solution of the subsidiary problem of minimization under consideration.

We now demonstrate that $\bar{\alpha}_1 > 0$.

Indeed, vector $\bar{p}_1$ satisfies condition (1.24), and $(a_j, \bar{p}_1) < 0$ according to lemma 1.4. Therefore

$$(a_i, \bar{p}_1) \leq 0, \quad i \in \mathcal{J}'_0. \quad (1.25)$$

For $i \notin \mathcal{J}'_0$ by the method of choosing set $\mathcal{J}'_0$,

$$(a_i, x_0) - b_i < 0.$$

Therefore

$$\bar{\alpha}_1 = \min_i \frac{b_i - (a_i, x_0)}{(a_i, \bar{p}_1)} > 0$$

since the minimum is taken only over those i for which $(a_i, \bar{p}_1) > 0$ and consequently practically over a certain subset of indices i that does not intersect $\mathcal{J}'_0$ according to (1.25). And

$$b_i - (a_i, x_0) > 0.$$

with such i .

The fact that $\bar{\alpha}_1 > 0$ indicates that all of the points $x_0 + \alpha p_1$ with $0 \leq \alpha \leq \bar{\alpha}_1$ satisfy conditions (1.2). Indeed, for $i \in \mathcal{Y}_0$

$$\begin{aligned} (a_i, x_0 + \alpha \bar{p}_1) - b_i &= (a_i, x_0) - b_i + \alpha (a_i, \bar{p}_1) \\ &= \alpha (a_i, \bar{p}_1) \begin{cases} = 0, & i \in \mathcal{Y}'_0, \\ < 0, & i = j. \end{cases} \end{aligned}$$

For $i \notin \mathcal{Y}_0$

$$(a_i, x_0 + \alpha \bar{p}_1) - b_i = (a_i, x_0) - b_i + \alpha (a_i, \bar{p}_1) < 0,$$

if $(a_i, p_1) \leq 0$; however if $(a_i, \bar{p}_1) > 0$, then

$$\alpha \leq \bar{\alpha}_1 \leq \frac{b_i - (a_i, x_0)}{(a_i, \bar{p}_1)}$$

and therefore

$$(a_i, x_0 + \alpha \bar{p}_1) - b_i \leq (a_i, x_0) - b_i + \frac{b_i - (a_i, x_0)}{(a_i, \bar{p}_1)} (a_i, \bar{p}_1) = 0.$$

Note that the sign of inequality in the last expression is to be considered strict if $\alpha < \bar{\alpha}_1$ or $\bar{\alpha}_1 < \frac{b_i - (a_i, x_0)}{(a_i, \bar{p}_1)}$.

According to the algorithm, two cases are possible: $\alpha_1 < \bar{\alpha}_1$ and $\alpha_1 \geq \bar{\alpha}_1$. In the first case we obtain a new point $x_1 = x_0 + \alpha_1 p_1$ that satisfies the relations

$$(a_i, x_1) - b_i = 0, \quad i \in \mathcal{Y}'_0, \quad (a_i, x_1) - b_i < 0, \quad i \notin \mathcal{Y}'_0. \quad (1.26)$$

In the second case, we obtain point $x_1 = x_0 + \bar{\alpha}_1 p_1$ and it is taken as a new initial point from which the algorithm begins to operate in checking case (1) or (2). Point x_1 satisfies conditions $(a_i, x_1) - b_i = 0$, $i \in \mathcal{Y}'_0$ and moreover equalities $(a_i, x_1) - b_i = 0$ with all $i \notin \mathcal{Y}'_0$ such that

$$\frac{b_i - (a_i, x_0)}{(a_i, \bar{p}_1)} = \bar{\alpha}_1.$$

Thus $\mathcal{Y}(x_1) \supset \mathcal{Y}'_0$, the inclusion being strict.

We return now to the case when $\alpha_1 < \bar{\alpha}_1$. Here the application of the method of conjugate gradients goes on, and so long as $\alpha_{k+1} < \bar{\alpha}_{k+1}$ holds, all of the points x_{k+1} continue to satisfy the relations (1.26) (like x_1), since by (1.14) and (1.9) with $P = P_{\mathcal{Y}'_0}$ we have

$$A_{\mathcal{Y}'_0} p_k = A_{\mathcal{Y}'_0} (I - P_{\mathcal{Y}'_0}) p_k = 0,$$

i.e. if written in the component form

$$(a_i, p_k) = 0, \quad i \in \mathcal{Y}'_0.$$

The inequalities

$$(a_i, x_k) - b_i < 0, \quad i \in \mathcal{Y}'_0$$

also will not be violated, for their violation would mean that the case where $\alpha_{k+1} \geq \bar{\alpha}_{k+1}$ takes place.

Thus, we have demonstrated that in case (1) the iterative process constructs successively points $x_0, x_1, \dots, x_k, k \geq 1$ and the value of $f(x)$ strictly decreases along this sequence because it is constructed by the method of conjugate directions. The last point x_k is either the minimum of $f(x)$ with constraints (1.16) or at this point $\mathcal{Y}(x_k)$ comprises strictly set $\mathcal{Y}'_0$.

In case (2) the direction of motion from point x_0 coincides with vector $p_1 = -(I - P_{\mathcal{Y}_0}) f'(x_0) \neq 0, (a_i, p_1) = 0, i \in \mathcal{Y}_0 (= \mathcal{Y}(x_0))$ and therefore it can be easily shown that $\bar{\alpha}_1 > 0$ and the method of conjugate directions makes it possible to make at least one nonzero step to the new point x_1 at which the value of $f(x)$ is strictly smaller.

All the proofs in this case are analogous to those given above. We obtain as a result the sequence of points $x_0, x_1, \dots, x_k, k \geq 1$, and x_k is either the minimum of $f(x)$ with constraints (1.18), or $\mathcal{Y}(x_k) \supset \mathcal{Y}_0$.

Note that in case (1) as well as in case (2) if point x_k is the minimum point of $f(x)$ with constraints (1.16), then in both cases x_k is the minimum of $f(x)$ on that face of the polyhedral set which is determined by expressions

$$(a_i, x) - b_i = 0, \quad i \in \mathcal{Y}(x_k) \quad (1.27)$$

since, by construction, $\mathcal{Y}(x_k) \supset \mathcal{Y}'_0$ in case (1) and $\mathcal{Y}(x_k) \supset \mathcal{Y}_0$ in case (2) and the minimum point on the broader set is the minimum point also on the narrower one.

We show now that after a finite number of steps starting with point x_0 we shall inevitably come to point x_k which is itself the minimum point of $f(x)$ with constraints (1.27). Indeed, it can be seen from the foregoing that if the method of conjugate gradients does not result in the finding of the minimum point, then it follows immediately that the set of i indices is extended and for them the next point obtained satisfies the relations $(a_i, x_k) - b_i = 0$. Since, by assumption, vectors $a_i, i \in \mathcal{Y}(x_k)$ are linearly independent, it is clear that this extension must be truncated after a finite number of steps not exceeding n , where n is the dimension of x .

Thus after a number of steps not exceeding n the algorithm described constructs the next point x_k which is the minimum of $f(x)$ with constraints (1.27).

Note that the sets $\mathcal{Y}(x_k)$ with different x_k are different for the value of function $f(x)$ decreases monotonically along the sequence con-

structed. Indeed, let x_m and x_k , $m < k$ be minimum points of $f(x)$ with constraints

$$(a_i, x) - b_i = 0,$$

$i \in \mathcal{Y}(x_m)$ and $i \in \mathcal{Y}(x_k)$ respectively. If $\mathcal{Y}(x_m) = \mathcal{Y}(x_k)$, then clearly $f(x_m) = f(x_k)$. But according to the construction of the process, $f(x_m) < f(x_k)$ with $m > k$ and thus the equality $\mathcal{Y}(x_m) = \mathcal{Y}(x_k)$ does not hold true.

On the other hand, all sets $\mathcal{Y}(x_k)$ are subsets of a finite set $\mathcal{Y} = \mathcal{Y}^- \cup \mathcal{Y}^0$ and therefore the number of such subsets is limited. It follows that the process proposed must be truncated after a finite number of steps. But this can occur only if the minimum point of $f(x)$ with constraints (1.2) has been found, otherwise as shown above the process can go on.

This proves that the process converges after a finite number of steps.

Remark. If matrix C is singular, then according to the theory of methods of conjugate directions it can occur that $(f'(x_k), p_{k+1}) \neq 0$ at point x_k but $(p_{k+1}, Cp_{k+1}) = 0$. In this case α_{k+1} cannot be calculated since

$$\alpha_{k+1} = - \frac{(f'(x_k), p_{k+1})}{(p_{k+1}, Cp_{k+1})}.$$

However in this case, $f(x_k + \alpha p_{k+1})$ decreases with increasing α and therefore we can assume that $\alpha_{k+1} = +\infty$ and perform the calculations as usual. If $\alpha_{k+1} < +\infty$, then the application of the method of conjugate gradients will terminate at point $x_{k+1} = x_k + \alpha_{k+1} p_{k+1}$; this violates the above argument in nothing. However, if α_{k+1} also has no limit, i.e. if $(a_i, p_{k+1}) \leq 0$ for all i , then the motion along the ray $x_k + \alpha p_{k+1}$ results in a decrease, without limit, of function $f(x)$. This means that the problem of quadratic programming which we consider has no solution since the lower bound of $f(x)$ with constraints (1.2) is $-\infty$.

Computational Aspects

The algorithm proposed comprises in essence only one complicated computational operation: the projecting of the gradient on a subspace, i.e. the calculation of the quantity $(I - P_{\mathcal{Y}})f'(x)$. There are two ways of performing this calculation.

The first one consists in a direct calculation of matrix $P_{\mathcal{Y}}$, i.e. $P_{\mathcal{Y}} = A_{\mathcal{Y}}^* (A_{\mathcal{Y}} A_{\mathcal{Y}}^*)^{-1} A_{\mathcal{Y}}$. This involves calculating matrix $(A_{\mathcal{Y}} A_{\mathcal{Y}}^*)^{-1}$. If this matrix is known, then the calculation of the required vector $u = -(A_{\mathcal{Y}} A_{\mathcal{Y}}^*)^{-1} A_{\mathcal{Y}} f'(x)$ is reduced to multiplying the matrix by the vector.

In order to diminish the amount of computations at each step, when set $\mathcal{Y}$ changes, one can make use of the fact that by deleting index j we delete in matrix $A_{\mathcal{Y}}A_{\mathcal{Y}}^*$ one column and one row, thus obtaining matrix $A_{\mathcal{Y}'}A_{\mathcal{Y}'}^*$. Just in the same way, in adding an additional index to set $\mathcal{Y}$, matrix $A_{\mathcal{Y}}A_{\mathcal{Y}}^*$ acquires an additional column and row.

This makes it possible to use the following recursive formulas known from linear algebra (see D. K. Faddeev and V. N. Faddeeva).

Let B be an arbitrary symmetric $n \times n$ matrix which can be written in the form

$$B = \begin{pmatrix} D & u \\ u^* & b \end{pmatrix}$$

where D is an $(n-1) \times (n-1)$ matrix, u is an $(n-1)$ -dimensional column-vector, u^* is its transposed vector, b is a number. Then it can be easily ascertained that

$$B^{-1} = \begin{pmatrix} D^{-1} + \frac{D^{-1}uu^*D^{-1}}{\alpha}, & -\frac{D^{-1}u}{\alpha} \\ -\frac{u^*D^{-1}}{\alpha}, & \frac{1}{\alpha} \end{pmatrix}$$

$$\alpha = b - u^*D^{-1}u.$$

Thus if matrix D^{-1} is known, then matrix B^{-1} , where B is obtained by adding the last column and the last row, can be obtained by simple calculations.

Conversely, if matrix B^{-1} is of the form

$$B^{-1} = \begin{pmatrix} G & p \\ p^* & m \end{pmatrix},$$

then for matrix D^{-1} we have

$$D^{-1} = G - \frac{pp^*}{m}.$$

Thus if the new matrix is obtained from the original one by deleting the last row and the last column or by adding a row and a column, then the inverse matrices are obtained by simple arithmetic operations. The fact that in the formulas given above we deleted the last column and row does not matter, for it can be easily checked that the transposition of rows in the original matrix leads simply to a transposition of columns in the inverted matrix, and the transposition of columns—to the transposition of rows.

Thus we have shown that the calculation of the matrix of projection can be performed by recursive formulas. The drawback of these recursive computations is that they may lead to a great cumulative computation error.

Let us describe another way of computation.

It was shown in the subsection on p. 151 that vector $p_0 = -(I - P_{\mathcal{Y}}) f'(x)$ is the solution of the problem of minimizing $(f'(x), p) + \frac{1}{2} \|p\|^2$ with constraints $A_{\mathcal{Y}} p = 0$. It is expedient to go over to the dual problem which, as demonstrated above, consists in the maximization of the quadratic function

$$-\frac{1}{2} \|f'(x) + A_{\mathcal{Y}}^* u\|^2$$

along vector u without constraints. This problem can be easily solved by the method of conjugate directions. As was shown in the subsection on p. 151 its solution is vector $u_0 = -(A_{\mathcal{Y}} A_{\mathcal{Y}}^*)^{-1} A_{\mathcal{Y}} f'(x)$, i.e. the vector which is required for the application of the algorithm for solving the general problem of quadratic programming. Vector p_0 is easily calculated using u_0 and the following formula:

$$\begin{aligned} p_0 &= -(I - P_{\mathcal{Y}}) f'(x) = -[f'(x) - A_{\mathcal{Y}}^* (A_{\mathcal{Y}} A_{\mathcal{Y}}^*)^{-1} A_{\mathcal{Y}} f'(x)] \\ &= -[f'(x) + A_{\mathcal{Y}}^* u_0], \end{aligned}$$

i.e.

$$p_0 = -[f'(x) + A_{\mathcal{Y}}^* u_0].$$

Thus in using the second way of computing, the operation is reduced to applying many times the standard procedure of the method of conjugate directions.

Problem of Quadratic Programming with Simple Constraints

The problem with simple constraints is understood to be the problem of minimizing

$$f(x) = \frac{1}{2} (x, Cx) + (d, x)$$

with constraints $x^i \geq 0$, $i \in \mathcal{J}$, where $\mathcal{J}$ is a subset of the set $\{1, 2, \dots, n\}$. In this case the algorithm of the subsection on p. 151 is considerably simplified. Instead of performing these simplifications formally, we shall formulate an algorithm for solving the problem. Its description will make it clear that the proof of its convergence after a finite number of steps coincides with the proof of the algorithm of the subsection on p. 151. So, let x_0 be an arbitrary point which satisfies constraints $x_0^i \geq 0$, $i \in \mathcal{J}$. Suppose that

$$\mathcal{J}(x) = \{i: x_i = 0, \quad i \in \mathcal{J}\}.$$

We shall describe now the procedure for one iteration with the initial point x_0 . We calculate the set $\mathcal{Y}(x_0)$. Two cases are possible.

(1) $(f'(x_0))^i = 0$, $i \in \mathcal{Y}(x_0)$, where $(f'(x_0))^i$ is the i -th component of vector $f'(x_0)$.

In this case point x_0 is the minimum point of $f(x)$ with constraints $x^i = 0$, $i \in \mathcal{Y}(x_0)$. If at the same time $(f'(x_0))^i \geq 0$ with $i \in \mathcal{Y}(x_0)$, then x_0 is the solution of the problem, for at point x_0 the necessary and sufficient conditions for a minimum are satisfied (see Chap. I, Sec. 3).

Let now $(f'(x_0))^i < 0$ for some $i \in \mathcal{Y}(x_0)$, and set

$$\mathcal{Y}' = \{i \in \mathcal{Y}(x_0): (f'(x_0))^i \geq 0\}.$$

We apply the method of conjugate gradients to minimize $f(x)$ taking as variables only x^i , $i \in \mathcal{Y}'$ and taking all the x^i for $i \in \mathcal{Y}'$ to be zero. The method of conjugate gradients requires the quantity $\bar{\alpha}_{k+1}$ be computed

$$\bar{\alpha}_{k+1} = \min_i \left(-\frac{x_k^i}{p_{k+1}^i} \right)$$

where the minimum is taken over all $i \in \mathcal{Y}'$ for which $p_{k+1}^i < 0$. Then we compare α_{k+1} and $\bar{\alpha}_{k+1}$.

If $\alpha_{k+1} < \bar{\alpha}_{k+1}$, then $x_{k+1}^i = x_k^i + \alpha_{k+1} p_{k+1}^i$, $i \in \mathcal{Y}'$, $x_{k+1}^i = x_k^i = 0$, $i \in \mathcal{Y}'$. If $\alpha_{k+1} \geq \bar{\alpha}_{k+1}$, then $x_{k+1}^i = x_k^i + \bar{\alpha}_{k+1} p_{k+1}^i$, $i \in \mathcal{Y}'$, $x_{k+1}^i = x_k^i = 0$, $i \in \mathcal{Y}'$. The calculation process will be truncated after a finite number of steps and a point x_{k+1} will be found such that it provides minimum of $f(x)$ subject to $x^i = 0$, $i \in \mathcal{Y}'$ or such that $\alpha_{k+1} \geq \bar{\alpha}_{k+1}$. Here, $\mathcal{Y}(x_k) \supset \mathcal{Y}'$ and the inclusion is such that there are $i \in \mathcal{Y}(x_k)$, but $i \notin \mathcal{Y}'$. In both cases point x_{k+1} is taken as the initial point and the process is repeated.

(2) There are indices i such that $(f'(x_0))^i \neq 0$, $i \in \mathcal{Y}(x_0)$. In this case we apply the method of conjugate gradients to minimize $f(x)$ with variables x^i , $i \in \mathcal{Y}(x_0)$. The components x^i , $i \in \mathcal{Y}(x_0)$ all the time remain equal to zero. Moreover as in case (1), at every step we calculate the quantity

$$\bar{\alpha}_{k+1} = \min_i \left(-\frac{x_k^i}{p_{k+1}^i} \right)$$

where the minimum is taken over all $i \in \mathcal{Y}(x_0)$, $p_{k+1}^i < 0$. The process stops in the same way as in case (1).

It is easily seen that an argument analogous to that given in the subsection on p. 151 results either in the proof of the convergence of the algorithm after a finite number of steps or in establishing the fact that $f(x)$ has no lower bound with conditions $x^i \geq 0$, $i \in \mathcal{Y}$.

2. METHOD OF FEASIBLE DIRECTIONS

The method of feasible directions was one of the first methods suggested for solving the problem of convex programming.

Suppose it is required to minimize function $f_0(x)$ with the constraints:

$$f_i(x) \leq 0, \quad i = 1, \dots, m, \quad Ax - b = 0 \quad (2.1)$$

where $x \in E^n$, $f_i(x)$, $i = 0, 1, \dots, m$ are convex continuously differentiable functions, A is an $l \times m$ matrix, b is an l -dimensional vector. Moreover, suppose that the gradients of functions $f_i(x)$, $i = 0, 1, \dots, m$ satisfy Lipschitz' condition:

$$\|f'_i(x_1) - f'_i(x_2)\| \leq C\|x_1 - x_2\| \quad (2.2)$$

and $\|f'_i(x)\| \leq K$ for all points x which are considered in what follows. We denote by D the admissible region, i.e. the set

$$D = \{x: f_i(x) \leq 0, \quad i = 1, \dots, m, \quad Ax - b = 0\}.$$

We shall assume in what follows that set D is compact and the condition of the gradients having bounds is fulfilled. Let x_0 be a point of D . We find a direction $p \in E^n$ such that with small α $(x_0 + \alpha p) \in D$, and, besides, $f_0(x_0 + \alpha p) < f_0(x_0)$. Such a direction is called feasible. Moving along this direction by one step α_1 we obtain a new point $x_1 = x_0 + \alpha_1 p \in D$. We take this point as the initial point and the process is repeated. The problem consists now in working out an effective method of finding feasible directions and choosing step α so as to provide for convergence to the minimum point.

Below we assume always that the following condition of nondegeneracy is fulfilled: there is a point $\bar{x}$ such that

$$A\bar{x} - b = 0, \quad f_i(\bar{x}) < 0, \quad i = 1, \dots, m.$$

Method of Choosing Feasible Directions

Let

$$\mathcal{J}_{\delta}^-(x) = \{i: f_i(x) \geq -\delta, \quad i = 1, \dots, m\}$$

for each point $x \in D$. Let $\xi_i > 0$, $i = 0, 1, \dots, m$ be arbitrary numbers. Consider the following problem at each point $x \in D$:

$$\begin{aligned} & \min \eta, \\ & (f'_i(x), p) \leq \xi_i \eta, \quad i \in \mathcal{J}_{\delta}^-(x) \cup \{0\}, \\ & Ap = 0, \quad \|p\| \leq 1 \end{aligned} \quad (2.3)$$

where η is a number and $\|p\|$ an arbitrary norm. To make (2.3) a problem of linear programming it is convenient to take as a norm

$$\|p\| = \max_{1 \leq i \leq n} |p^i|.$$

Let $p_\delta(x)$, $\eta_\delta(x)$ be a solution of problem (2.3). Since vector $p = 0$, $\eta = 0$ satisfies constraints (2.3), clearly we have $\eta_\delta(x) \leq 0$. We demonstrate that $p_\delta(x)$ is a feasible direction if $\eta_\delta(x) < 0$. Indeed, let $\alpha > 0$. For $i = 0$ we have by Taylor's formula

$$\begin{aligned} f_0(x + \alpha p_\delta(x)) &= f_0(x) + \alpha (f'_0(\theta_0), p_\delta(x)) \\ &= f_0(x) + \alpha (f'_0(x), p_\delta(x)) + \alpha (f'_0(\theta_0) - f'_0(x), p_\delta(x)) \\ &\leq f_0(x) + \alpha (f'_0(x), p_\delta(x)) + \alpha^2 C \|p_\delta(x)\|^2 \end{aligned}$$

where $\theta_0 = x + \xi_0 \alpha p_\delta(x)$, $0 \leq \xi_0 \leq 1$ and where we have used the fact that

$$\|f'_0(\theta_0) - f'_0(x)\| \leq C \|\theta_0 - x\| \leq C \alpha \|p_\delta(x)\|.$$

Further by (2.3), $(f'_0(x), p_\delta(x)) \leq \xi_0 \eta_\delta(x)$. Therefore

$$f_0(x + \alpha p_\delta(x)) \leq f_0(x) + \alpha \xi_0 \eta_\delta(x) + \alpha^2 C \|p_\delta(x)\|^2. \quad (2.4)$$

By analogy, for $i \in \mathcal{J}_\delta^-(x)$

$$f_i(x + \alpha p_\delta(x)) \leq f_i(x) + \alpha \xi_i \eta_\delta(x) + \alpha^2 C \|p_\delta(x)\|^2. \quad (2.5)$$

Further, for $i \in \mathcal{J}_\delta^+(x)$

$$f_i(x + \alpha p_\delta(x)) = f_i(x) + \alpha (f'_i(\theta_i), p_\delta(x)) \leq f_i(x) + \alpha K \|p_\delta(x)\|. \quad (2.6)$$

We now choose $\alpha > 0$ such as to satisfy the inequalities

$$\begin{aligned} f_0(x + \alpha p_\delta(x)) &\leq f_0(x) + \frac{1}{2} \alpha \xi_0 \eta_\delta(x), \\ f_i(x + \alpha p_\delta(x)) &\leq 0, \quad i \in \mathcal{J}_\delta^-(x), \quad f_i(x + \alpha p_\delta(x)) \leq 0, \quad i \in \mathcal{J}_\delta^+(x). \end{aligned} \quad (2.7)$$

To satisfy these inequalities it is sufficient that

$$\begin{aligned} 1 + \frac{\alpha C \|p_\delta(x)\|^2}{\xi_0 \eta_\delta(x)} &\leq \frac{1}{2}, \quad \xi_i \eta_\delta(x) + \alpha C \|p_\delta(x)\|^2 \leq 0, \quad i \in \mathcal{J}_\delta^-(x), \\ -\delta + \alpha K \|p_\delta(x)\| &\leq 0, \quad i \in \mathcal{J}_\delta^+(x) \end{aligned} \quad (2.8)$$

be true since, by (2.4), (2.5), the following inequalities hold:

$$\begin{aligned} f_0(x + \alpha p_\delta(x)) &\leq f_0(x) + \alpha \xi_0 \eta_\delta(x) \left[1 + \frac{\alpha C \|p_\delta(x)\|^2}{\xi_0 \eta_\delta(x)} \right], \\ f_i(x + \alpha p_\delta(x)) &\leq f_i(x) + \alpha [\xi_i \eta_\delta(x) + \alpha C \|p_\delta(x)\|^2], \quad i \in \mathcal{J}_\delta^-(x), \end{aligned}$$

and since $f_i(x) \leq -\delta$ for $i \in \mathcal{J}_\delta(x)$, then

$$f_i(x + \alpha p_\delta(x)) \leq -\delta + \alpha K \|p_\delta(x)\|.$$

From inequalities (2.8) we obtain:

$$\alpha \leq -\frac{1}{2} \frac{\xi_0 \eta_\delta(x)}{C \|p_\delta(x)\|^2}, \quad \alpha \leq -\frac{\xi_i \eta_\delta(x)}{C \|p_\delta(x)\|^2}, \quad \alpha \leq \frac{\delta}{K \|p_\delta(x)\|}. \quad (2.9)$$

Thus if α satisfies inequalities (2.9), then inequalities (2.7) are fulfilled and it follows that $p_\delta(x)$ is really a feasible direction for $Ap_\delta(x) = 0$ and therefore $A(x + \alpha p_\delta(x)) - b = Ax - b + \alpha Ap_\delta(x) = 0$. We now show that if point x does not coincide with the solution of the problem x_* which is the minimum point of $f_0(x)$ in region D , then $\eta_\delta(x) < 0$ with any sufficiently small δ .

Lemma 2.1. *Let $x \in D$ be not a solution of the problem of minimization of $f_0(x)$ with constraints (2.3). Then $\eta_\delta(x) < 0$ for any sufficiently small δ .*

Proof. Recall that we assume the conditions of nondegeneracy to be fulfilled, i.e. there is a point $\bar{x}$ such that

$$A\bar{x} - b = 0, \quad f_i(\bar{x}) \leq \sigma, \quad i = 1, \dots, m, \quad \sigma < 0. \quad (2.10)$$

Let point x_* be the solution of the problem. Further, if

$$\mathcal{J}_0^-(x) = \{i: f_i(x) = 0, \quad i = 1, \dots, m\},$$

then with $\delta < \delta_0$

$$-\delta_0 = \max_{i \in \mathcal{J}_0^-(x)} f_i(x),$$

$\mathcal{J}_\delta^-(x) = \mathcal{J}_0^-(x)$. Indeed, if $i \in \mathcal{J}_\delta^-(x)$, $\delta < \delta_0$, then $f_i(x) \geq -\delta$. But for all $i \in \mathcal{J}_0^-(x)$, $f_i(x) \leq -\delta_0 < -\delta$, i.e. $i \in \mathcal{J}_\delta^-(x)$. Suppose that $\delta < \delta_0$ and so $\mathcal{J}_\delta^-(x) = \mathcal{J}_0^-(x)$. We set

$$x_\rho = \rho \bar{x} + (1 - \rho) x_*, \quad 0 < \rho \leq 1.$$

Then due to the convexity of functions $f_i(x)$, $i = 0, 1, \dots, m$ and the fact that $f_i(x_*) \leq 0$, $i = 1, \dots, m$, we obtain

$$f_i(x_\rho) \leq \rho f_i(\bar{x}) + (1 - \rho) f_i(x_*) \leq \rho \sigma, \quad i = 1, \dots, m.$$

Further, for $i \in \mathcal{J}_\delta^-(x)$, $f_i(x) = 0$ and therefore for $0 < \lambda < 1$,

$$\begin{aligned} \lambda \rho \sigma &\geq \lambda f_i(x_\rho) = \lambda f_i(x_\rho) + (1 - \lambda) f_i(x) \geq f_i(\lambda x_\rho + (1 - \lambda) x) \\ &= f_i(x + \lambda(x_\rho - x)) - f_i(x) \geq \lambda (f'_i(x), x_\rho - x) \end{aligned}$$

where we have used the inequality (Chap. I, Sec. 2)

$$f(y) - f(x) \geq (f'(x), y - x)$$

which holds for any differentiable convex function.

Thus

$$\rho\sigma \geq (f'_i(x), x_\rho - x), \quad i \in \mathcal{J}_\delta^-(x). \quad (2.11)$$

Further, since point x does not provide a minimum of $f(x)$ in D

$$0 > \gamma = f_0(x_*) - f_0(x) \geq (f'_0(x), x_* - x).$$

Hence

$$\begin{aligned} (f'_0(x), x_\rho - x) &= \rho (f'_0(x), \bar{x} - x) + (1 - \rho) (f'_0(x), x_* - x) \\ &\leq \rho (f'_0(x), \bar{x} - x) + (1 - \rho) \gamma. \end{aligned} \quad (2.12)$$

It follows from (2.11) and (2.12) that with sufficiently small $\rho > 0$ the following inequalities are satisfied:

$$(f'_0(x), \bar{p}_\rho) < 0, \quad (f'_i(x), \bar{p}_\rho) < 0, \quad i \in \mathcal{J}_\delta^-(x) \quad (2.13)$$

where $\bar{p}_\rho = x_\rho - x$, and the fact that $\sigma < 0$, $\gamma < 0$ has been taken into account. Take $p_\rho = \bar{p}_\rho$ if $\|\bar{p}_\rho\| \leq 1$, and $p_\rho = \frac{\bar{p}_\rho}{\|\bar{p}_\rho\|}$ if $\|\bar{p}_\rho\| > 1$, and so $\|p_\rho\| \leq 1$. Besides, we take

$$\eta_\rho = \max_{i \in \mathcal{J}_\delta^-(x) \cup \{0\}} \frac{(f'_i(x), p_\rho)}{\xi_i}.$$

By (2.13), $\eta_\rho < 0$ and the inequalities

$$(f'_i(x), p_\rho) \leq \xi_i \eta_\rho, \quad i \in \mathcal{J}_\delta^-(x) \cup \{0\}, \quad \eta_\rho < 0 \quad (2.14)$$

are satisfied. Further, since $x_\rho = \rho \bar{x} + (1 - \rho) x_*$ and the equalities

$$A\bar{x} - b = 0, \quad Ax_* - b = 0$$

hold, then

$$Ax_\rho - b = 0.$$

Note that $p_\rho = \alpha \bar{p}_\rho = \alpha (x_\rho - x)$, where $0 < \alpha \leq 1$. Therefore

$$Ap_\rho + (Ax - b) = \alpha [Ax_\rho - b] + (1 - \alpha) [Ax - b] = 0. \quad (2.15)$$

It follows from (2.14) and (2.15) that vector p_ρ and the quantity η_ρ satisfy conditions (2.3). Since $\eta_\rho < 0$, so much the more $\eta_\delta(x) < 0$, for $\eta_\delta(x) \leq \eta_\rho$ by definition. The lemma is proved.

Algorithm of Method of Feasible Directions

Let $x_0 \in D$ be an arbitrary first approximation and $\delta_0 > 0$. We describe the general step of the algorithm. Let point $x_k \in D$ be obtained at the k -th step and $\delta_k > 0$.

Having solved the problem

$$\begin{aligned} \min \eta, \\ (f'_i(x_k), p) \leq \xi_i \eta, \quad i \in \mathcal{J}_\delta^-(x_k) \cup \{0\}, \\ Ap = 0, \quad \|p\| \leq 1, \end{aligned}$$

we obtain $p_{\delta_k}(x_k) \equiv p_k$ and $\eta_{\delta_k}(x_k) \equiv \eta_k$.

Remark. If we take the quantity $\max_i |p^i|$ as the norm of vector p , then the above problem is a problem of linear programming and can be solved by one of the standard methods.

Two cases are possible.

(1) $\eta_k < -\delta_k$. We take successively $\alpha = \frac{1}{2^i}$, $i = 0, 1, \dots$, and find the first i_0 such that the following inequalities

$$\begin{aligned} f_0\left(x_k + \frac{1}{2^{i_0}} p_k\right) &\leq f_0(x_k) + \frac{1}{2 \cdot 2^{i_0}} \xi_0 \eta_k, \\ f_i\left(x_k + \frac{1}{2^{i_0}} p_k\right) &\leq 0, \quad i = 1, \dots, m \end{aligned}$$

are satisfied. We take $\alpha_k = \frac{1}{2^{i_0}}$ and $x_{k+1} = x_k + \alpha_k p_k$, $\delta_{k+1} = \delta_k$, so that

$$\begin{aligned} f_0(x_{k+1}) &\leq f_0(x_k) + \frac{1}{2} \alpha_k \xi_0 \eta_k, \\ f_i(x_{k+1}) &\leq 0, \quad i = 1, \dots, m. \end{aligned} \tag{2.16}$$

(2) $\eta_k \geq -\delta_k$. We take

$$x_{k+1} = x_k, \quad \delta_{k+1} = \frac{1}{2} \delta_k.$$

Thus in the first case there is a shift to a new point, in the second one there is no such a shift.

We now formulate also the condition of the halt of the algorithm: if at a certain step k , $\delta_k < \delta^0(x_k)$, where

$$\delta^0(x_k) = - \max_{i \in \bar{\mathcal{J}}_0^-(x_k)} f_i(x_k)$$

and $\eta_k = 0$, then x_k is the solution of the problem set above, i.e. x_k is the minimum point of $f_0(x)$ with constraints (2.1).

Substantiation of Convergence of the Algorithm

We show that if the sequence $\{x_k\}$ is truncated at a certain step k because the conditions of the halt have been fulfilled, then x_k is really the solution of the problem. Indeed, let the conditions of the halt be fulfilled, i.e. $\eta_k = \eta_{\delta_k}(x_k) = 0$, and

$$\delta_k < \delta^0(x_k) = - \max_{i \in \bar{\mathcal{J}}_0^-(x_k)} f_i(x_k). \quad (2.17)$$

But as was shown in proving lemma 2.1, if conditions (2.17) are fulfilled, then $\eta_{\delta_k}(x_k) < 0$ provided x_k is not the solution of the problem. But it was assumed that $\eta_k = 0$ so it follows that x_k is the minimum point of $f_0(x)$ with $x \in D$. Let the iterative process be now continued without limit so that we have an infinite sequence $\{x_k\}$, $k = 0, 1, \dots$. Let x_k be a point at which $\eta_k < -\delta_k$, i.e. case (1) takes place. Then making use of estimates (2.9) and the fact that $\|p_k\| = \|p_{\delta_k}(x_k)\| \leq 1$, we can state that if

$$\begin{aligned} \alpha &\leq -\frac{1}{2} \frac{\xi_0 \eta_k}{C}, \\ \alpha &\leq -\frac{\xi_i \eta_k}{C}, \quad i \in \mathcal{J}_k^-, \\ \alpha &\leq \frac{\delta_k}{K}, \quad i \in \bar{\mathcal{J}}_k^-, \end{aligned} \quad (2.18)$$

where to shorten the written form we set $\bar{\mathcal{J}}_k^- = \bar{\mathcal{J}}_{\delta_k}^-(x_k)$, then the inequalities (2.7) will hold:

$$\begin{aligned} f_0(x_k + \alpha p_k) &\leq f_0(x_k) + \frac{1}{2} \alpha \xi_0 \eta_k, \\ f_i(x_k + \alpha p_k) &\leq 0, \quad i = 1, \dots, m. \end{aligned}$$

Now, recall that according to the algorithm, quantity α_k coincides with the first of the quantities $\frac{1}{2^i}$, $i = 0, 1, \dots$, satisfying inequalities (2.16). It follows that after a finite number of trials these inequalities will be satisfied.

Let i_0 be the first of the indices satisfying the inequalities and so $\alpha = \frac{1}{2^{i_0}}$. This means that with $\alpha = \frac{1}{2^{i_0-1}}$ the inequalities (2.16) were not satisfied and therefore α did not satisfy (2.18), i.e.

$$\frac{1}{2^{i_0-1}} > \min \left\{ -\frac{1}{2} \frac{\xi_0 \eta_k}{C}, \frac{\delta_k}{K}, \min_{i \in \bar{\mathcal{J}}_k^-} -\frac{\xi_i \eta_k}{C} \right\}.$$

Therefore

$$\alpha_k = \frac{1}{2^{i_0}} > \frac{1}{2} \min \left\{ -\frac{\xi_0 \eta_k}{2C}, \frac{\delta_k}{K}, \min_{i \in \mathcal{J}_h^-} -\frac{\xi_i \eta_k}{C} \right\}. \quad (2.19)$$

If we take into account that in the case under consideration $-\eta_k > \delta_k$, then inequality (2.19) can be made stronger by substituting δ_k for $-\eta_k$. Then we obtain:

$$\alpha_k > \frac{\delta_k}{2} \varepsilon_0, \quad \varepsilon_0 = \min \left\{ \frac{\xi_0}{2C}, \frac{\xi_1}{C}, \frac{\xi_2}{C}, \dots, \frac{\xi_m}{C}, \frac{1}{K} \right\}. \quad (2.20)$$

Using now inequalities (2.16), (2.20) and the fact that $\eta_k < -\delta_k < 0$, we obtain

$$f_0(x_{k+1}) \leq f_0(x_k) + \frac{1}{2} \alpha_k \xi_0 \eta_k \leq f_0(x_k) - \frac{\xi_0 \varepsilon_0}{4} \delta_k^2. \quad (2.21)$$

It follows from this inequality that $\delta_k \rightarrow 0$ as $k \rightarrow \infty$. Indeed, as the sequence $\{\delta_k\}$, $k = 0, 1, \dots$, decreases monotonically, and if $\delta_{k+1} < \delta_k$, then $\delta_{k+1} = \frac{1}{2} \delta_k$, and the fact that δ_k does not tend to zero can mean only that $\delta_k = \delta > 0$ for all sufficiently large k . But if δ_k remains constant, then the condition $\eta_k < -\delta_k$ is fulfilled and thus the inequality (2.21) holds.

Thus for all sufficiently large k ($k \geq k_0$), $\delta_k = \delta$ and the inequality

$$f_0(x_{k+1}) \leq f_0(x_k) - \frac{\xi_0 \varepsilon_0}{4} \delta^2$$

is fulfilled.

Therefore

$$f_0(x_N) \leq f_0(x_{k_0}) - (N - k_0) \frac{\xi_0 \varepsilon_0}{4} \delta^2,$$

i.e. $f_0(x_N) \rightarrow -\infty$ as $N \rightarrow \infty$. But this contradicts the fact that the continuous function $f_0(x)$ in the compact region D is bounded.

Thus we have shown that $\delta_k \rightarrow 0$. This means that the initial δ_0 is successively halved an infinite number of times, i.e. that case (2) takes place an infinite number of times: $\eta_k \geq -\delta_k$.

Let $\mathcal{Y}$ be a set of indices k for which case (2) took place. Then $\eta_k \rightarrow 0$ as $k \rightarrow \infty$, $k \in \mathcal{Y}$. This follows immediately from the inequality $-\delta_k \leq \eta_k \leq 0$ and from the fact that $\delta_k \rightarrow 0$. Consider the sequence of points $x_k \in D$, $k \in \mathcal{Y}$. As D is a compact set, one can assume, without loss of generality, that $\{x_k\}$ converges to a certain point x_* . We shall demonstrate now that x_* is the minimum point of $f_0(x)$ in D .

Suppose that the opposite is true, i.e. that point x_* is not the minimum point of $f_0(x)$ in D . Then on the basis of lemma 2.1, we can affirm that with all $\delta < \delta^0(x_*)$

$$\delta^0(x_*) = - \max_{i \in \mathcal{J}_0^-(x_*)} f_i(x_*),$$

$\mathcal{J}_\delta^-(x_*) = \mathcal{J}_0^-(x_*)$ and $\eta_\delta(x_*) < 0$. Moreover, since $\mathcal{J}_\delta^-(x_*) = \mathcal{J}_0^-(x_*)$, we have $\eta_\delta(x_*) = \eta_0(x_*) < 0$. Further, $\mathcal{J}_{\delta_k}^-(x_k) \subseteq \mathcal{J}_0^-(x_*)$ with sufficiently large $k \in \mathcal{Y}$. Indeed, suppose that $i \notin \mathcal{J}_0^-(x_*)$, then $f_i(x_*) < 0$. Therefore because of the fact that $\delta_k \rightarrow 0$, with sufficiently large k we have $f_i(x_*) < -\delta_k$, and since $x_k \rightarrow x_*$, with great k we also have, $f_i(x_k) < -\delta_k$, i.e. $i \notin \mathcal{J}_{\delta_k}^-(x_k)$. Thus if $i \notin \mathcal{J}_0^-(x_*)$, then with great k we have $i \notin \mathcal{J}_{\delta_k}^-(x_k)$, i.e. $\mathcal{J}_{\delta_k}^-(x_k) \subseteq \mathcal{J}_0^-(x_*)$. Since by assumption x_* is not the minimum point of $f_0(x)$ in D , there is a vector $p(x_*)$ such that $Ap(x_*) = 0$, $\|p(x_*)\| \leq 1$,

$$(f'_i(x_*), p(x_*)) \leq \xi_i \eta_0(x_*), \quad i \in \mathcal{J}_0^-(x_*) \cup \{0\}$$

and, as mentioned above, $\eta_0(x_*) < 0$. But then by continuity with great k the following relations hold

$$(f'_i(x_k), p(x_*)) \leq \frac{1}{2} \xi_i \eta_0(x_*), \quad i \in \mathcal{J}_k^- \cup \{0\},$$

$$Ap(x_*) = 0, \quad \|p(x_*)\| \leq 1,$$

for $x_k \rightarrow x_*$, $\mathcal{J}_k^- \subseteq \mathcal{J}_0^-(x_*)$. However, the last relations mean that

$$\eta_k \equiv \eta_{\delta_k}(x_k) \leq \frac{1}{2} \eta_0(x_*) < 0$$

with all sufficiently great k and this contradicts the fact that $\eta_k \rightarrow 0$, $k \rightarrow +\infty$, $k \in \mathcal{Y}$ which was proved above. The contradiction obtained proves that x_* is the minimum point of $f_0(x)$ in D .

Theorem 2.1. *The sequence of points $\{x_k\}$ constructed by the method of feasible directions has the property that $f_0(x_k)$ without increasing monotonically tends to $f_0(x_*)$, where x_* is the minimum point of $f_0(x)$ in region D .*

Proof. By construction of sequence $\{x_k\}$ we have $f_0(x_{k+1}) \leq f_0(x_k)$ and so this sequence of numbers does not increase monotonically. Since it has a lower bound, it converges to a certain limit, $\bar{f}_0$. However, it was shown above that there is a subsequence of $\{x_k\}$, $k \in \mathcal{Y}$ such that $x_k \rightarrow x_*$. Therefore $f_0(x_k) \rightarrow f(x_*)$. As the whole sequence converges to the same limit as the subsequence, it follows that $f_0(x_k) \rightarrow f_0(x_*)$. Q.E.D.

Remark 1. Among the constraints $f_i(x) \leq 0$ there can be such that for them functions $f_i(x)$ are linear. It is easy to show by a slight extension of the preceding argument that with these indices i we can take $\xi_i = 0$.

Besides, in this case the condition of nondegeneracy can also be weakened, viz. it suffices to require that there be a point $\bar{x} \in D$ such that $f_i(\bar{x}) < 0$ only for those indices i for which $f_i(x)$ are nonlinear.

Remark 2. Sequence $\{x_k\}$ itself, speaking generally, can lack convergence; however, if point x_* of the minimum of $f_0(x)$ with $x \in D$ is unique, it can easily be seen that $x_k \rightarrow x_*$. Unfortunately, the rate of convergence of the method of feasible directions is as yet unknown.

Construction of the Initial Approximation

The application of the method of feasible directions requires the knowledge of the initial approximation in region D . To obtain this initial approximation we can use the same method of feasible directions by applying it to the problem of minimization of number η with constraints

$$f_i(x) - \eta \leq 0, \quad i = 1, \dots, m, \quad Ax - b = 0. \quad (2.22)$$

As there is a point $\bar{x}$ such that

$$f_i(\bar{x}) < 0, \quad i = 1, \dots, m, \quad A\bar{x} - b = 0,$$

the minimum value of η with the constraints described is strictly less than zero and therefore after a finite number of steps we obtain point x and η such that $\eta < 0$ and the inequalities (2.22) will be satisfied. This means that the obtained point x satisfies the constraints of the original problem and can be taken as the initial point for applying the method of feasible directions.

3. METHOD OF CONDITIONAL GRADIENT AND NEWTON'S METHOD

The method of conditional gradient can be used for solving the problem of the minimization of a nonlinear function in a region in which the problem of the minimization of a linear function can be solved without great difficulties.

Suppose that $f(x)$, $x \in E^n$ is a continuously differentiable function in a compact convex region Ω , and the gradient $f'(x)$ of function $f(x)$ in Ω satisfies Lipschitz' condition, i.e.

$$\|f'(x_1) - f'(x_2)\| \leq L\|x_1 - x_2\| \quad (3.1)$$

for all the points of region Ω .

The method of conditional gradient consists in the following.

Let x_k , the approximation at the k -th step of the iterative process, be already constructed. Calculate $f'(x_k)$ and find the minimum point of linear function $(f'(x_k), z)$ in Ω . Let it be point $z(x_k)$. Take $p_k = z(x_k) - x_k$ and $x_{k+1} = x_k + \alpha_k p_k$, where $\alpha_k > 0$ is the length of the step in direction p_k . Point x_{k+1} is taken as the initial one and the procedure is then repeated.

It will be demonstrated below that with a definite rule of calculating α_k , the process converges and the bounds on the rate of convergence will be established. The same problems will be analyzed for Newton's method which differs from the method of conditional gradient in that the function being minimized is approximated at each iteration by a quadratic form (while in the method of conditional gradient the approximation is linear).

Rule for Choosing the Step Length

Let x be an arbitrary point in Ω . We denote by $z(x)$ a minimum point of function $(f'(x), z)$ in Ω such that

$$(f'(x), z(x)) \leq (f'(x), z), \quad z \in \Omega. \quad (3.2)$$

We take $p(x) = z(x) - x$,

$$\eta(x) = \min_{z \in \Omega} (f'(x), z - x) = (f'(x), p(x)).$$

By (3.2), $\eta(x) \leq 0$. We are interested in the estimate for the increase of the function value in moving from point x in direction $p(x)$. Using Taylor's formula and (3.1) we obtain:

$$\begin{aligned} f(x + \alpha p(x)) &= f(x) + \alpha (f'(\theta), p(x)) \\ &= f(x) + \alpha (f'(x), p(x)) + \alpha (f'(\theta) - f'(x), p(x)) \\ &\leq f(x) + \alpha \eta(x) + \alpha^2 L \|p(x)\|^2 \end{aligned}$$

where $\theta = x + \xi \alpha p(x)$, $0 \leq \xi \leq 1$. Thus

$$f(x + \alpha p(x)) \leq f(x) + \alpha (\eta(x) + \alpha L \|p(x)\|^2). \quad (3.3)$$

It follows directly from this formula that with

$$\alpha \leq -\frac{1}{2} \frac{\eta(x)}{L \|p(x)\|^2} \quad (3.4)$$

the following estimate holds:

$$f(x + \alpha p(x)) \leq f(x) + \frac{\alpha \eta(x)}{2}. \quad (3.5)$$

Description of the Algorithm

The algorithm begins with an arbitrary point x_0 of region Ω . We describe now the general step.

Let point x_k be already constructed, $k \geq 0$. Having solved the problem of minimization of $(f'(x_k), z)$ in Ω , calculate $z(x_k)$, $p(x_k)$, $\eta(x_k)$. Construct point $x_{k+1} = x_k + \alpha_k p(x_k)$, where α_k is taken to be 2^{-i_0} and i_0 is that of the indices $i = 0, 1, \dots$, which is the first to satisfy the inequality

$$f(x_k + 2^{-i} p(x_k)) \leq f(x_k) + 2^{-i} \frac{\eta(x_k)}{2}. \quad (3.6)$$

The condition of the halt: the process stops if $\eta(x_k) = 0$.

Substantiation of Convergence of the Algorithm and Estimation of Its Rate of Convergence

According to the just given rule of choosing the step length, the following inequality is fulfilled:

$$f(x_{k+1}) \leq f(x_k) + \frac{\alpha_k \eta(x_k)}{2}. \quad (3.7)$$

In order to substantiate the convergence of the algorithm, it is necessary first of all to demonstrate that inequalities (3.6), (3.7) can always be satisfied. In fact by (3.4) and (3.5), inequality (3.6) will be satisfied as soon as inequality

$$2^{-i} \leq -\frac{1}{2} \frac{\eta(x_k)}{L \|p(x_k)\|^2}$$

is satisfied, and since i_0 is the first index satisfying (3.6), we have

$$2\alpha_k = 2^{-(i_0-1)} > -\frac{1}{2} \frac{\eta(x_k)}{L \|p(x_k)\|^2}.$$

hence

$$\alpha_k > \frac{1}{4L} \frac{-\eta(x_k)}{\|p(x_k)\|^2}. \quad (3.8)$$

It follows from the foregoing that if $\eta(x_k) < 0$, then inequality (3.6) will be satisfied after a finite number of trials and the α_k chosen will satisfy inequality (3.8).

Lemma 3.1. *If $\{x_k\}$, $k = 0, 1, \dots$, is a sequence of points obtained in implementing the algorithm of the method of conditional gradient, then $x_k \in \Omega$, $f(x_k)$ decreases monotonically and $\eta(x_k) \rightarrow 0$ as $k \rightarrow +\infty$.*

Proof. Let $x_k \in \Omega$ for $k \leq m$. We show that $x_{m+1} \in \Omega$. Indeed, $0 \leq \alpha_k \leq 1$ and $z(x_k) \in \Omega$. Therefore

$$\begin{aligned} x_{m+1} &= x_m + \alpha_m p(x_m) = x_m + \alpha_m (z(x_m) - x_m) \\ &= (1 - \alpha_m) x_m + \alpha_m z(x_m) \in \Omega, \end{aligned}$$

for region Ω is convex.

Note now that $\|p(x_k)\|$ has a limit, a certain constant C , since $p(x_k) = z(x_k) - x_k$, $z(x_k) \in \Omega$, $x_k \in \Omega$ and Ω is a compact set.

Using now formulas (3.8), (3.7) we obtain

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{8LC^2} \eta^2(x_k). \quad (3.9)$$

Adding (3.9) for all $k = 0, 1, \dots, m-1$ we obtain

$$f(x_m) - f(x_0) \leq -\frac{1}{8LC^2} \sum_{k=0}^{m-1} \eta^2(x_k).$$

Since region Ω is compact and function $f(x)$ continuous, we have $f(x_m) \geq f_*$, where f_* is the minimum value of $f(x)$ in Ω . Therefore

$$\sum_{k=0}^{m-1} \eta^2(x_k) \leq 8LC^2 [f(x_0) - f(x_m)] \leq 8LC^2 [f(x_0) - f_*].$$

Hence it follows that the series

$$\sum_{k=0}^{\infty} \eta^2(x_k)$$

converges. This is possible only if $\eta(x_k) \rightarrow 0$. The lemma is proved.

It follows from the condition of the halt of an algorithm and lemma (3.1) that, in general, either the algorithm stops after a finite number of steps and the condition $\eta(x_k) = 0$ is fulfilled or a monotonically decreasing sequence of $f(x_k)$ values of function $f(x)$ is obtained.

In the first case, the condition $\eta(x_k) = 0$, by (3.2), is equivalent to the following one:

$$(f'(x_k), x_k) = (f'(x_k), z(x_k)) \leq (f'(x_k), z), \quad z \in \Omega.$$

The last expression is nothing else but the necessary condition for function $f(x)$ assuming its minimum value at point x_k (see Chap. I, Sec. 3).

The second case is the subject of the following lemma.

Lemma 3.2. *At any limit point of sequence $\{x_k\}$, $k = 0, 1, \dots$, the necessary conditions for the minimum of $f(x)$ in set Ω are fulfilled.*

Proof. Let x_* be a limit point of $\{x_k\}$, i.e. there is a subsequence $\{x_{k_j}\}$, $j \rightarrow \infty$, such that $x_{k_j} \rightarrow x_*$. The following relations hold:

$$\begin{aligned}\eta(x_{k_j}) &= (f'(x_{k_j}), z(x_{k_j}) - x_{k_j}), \\ (f'(x_{k_j}), z(x_{k_j})) &\leq (f'(x_{k_j}), z), \quad z \in \Omega.\end{aligned}$$

Without loss of generality, we can take that $z(x_{k_j}) \rightarrow z_*$. Since $\eta(x_k) \rightarrow 0$ and $f'(x)$ depends on x continuously, it follows from the above relations that

$$\begin{aligned}(f'(x_*), z_* - x_*) &= 0 \\ (f'(x_*), z_*) &\leq (f'(x_*), z), \quad z \in \Omega.\end{aligned}$$

Hence

$$(f'(x_*), x_*) \leq (f'(x_*), z), \quad z \in \Omega,$$

and this is the proof of the lemma.

Theorem 3.1. *Let function $f(x)$ be convex. Then*

$$\lim_{k \rightarrow \infty} f(x_k) = f_*$$

where $f_* = \min_{x \in \Omega} f(x)$. Moreover, the estimate

$$f(x_k) - f_* \leq \frac{C}{k}$$

(C is a positive constant) holds.

Proof. As $f(x)$ is a convex function, the following inequality holds:

$$f_* - f(x) = f(x_*) - f(x) \geq (f'(x), x_* - x) \geq \min_{z \in \Omega} (f'(x), z - x) = \eta(x).$$

Thus $0 \leq f(x) - f_* \leq -\eta(x)$. Therefore with all k

$$0 \leq f(x_k) - f_* \leq -\eta(x_k). \quad (3.10)$$

It follows from lemma 3.1 that $\eta(x_k) \rightarrow 0$. Therefore, the last inequality shows that $f(x_k) \rightarrow f_*$ and this proves the first part of the theorem.

Combining (3.9) and (3.10) we obtain

$$(f(x_{k+1}) - f_*) - (f(x_k) - f_*) \leq -\frac{1}{8LC^2} [f(x_k) - f_*]^2.$$

With the notation $\varphi_k = f(x_k) - f_*$ we obtain

$$\varphi_{k+1} \leq \varphi_k \left(1 - \frac{\varphi_k}{8LC^2} \right)$$

or

$$\varphi_{k+1} \leq \varphi_k (1 - \kappa \varphi_k), \quad \kappa = \frac{1}{8LC^2}.$$

Taking $\varphi_k = \frac{\gamma_k}{k}$ we obtain

$$\frac{\gamma_{k+1}}{\gamma_k} \leq \frac{k+1}{k} \left(1 - \kappa \frac{\gamma_k}{k} \right) \quad (3.11)$$

or

$$\frac{\gamma_{k+1}}{\gamma_k} \leq 1 + \frac{1}{k} - \frac{\kappa (k+1) \gamma_k}{k^2}.$$

With each k there are two possible cases:

$$(1) \quad \frac{\gamma_{k+1}}{\gamma_k} \leq 1, \text{ i. e. } \gamma_{k+1} \leq \gamma_k.$$

$$(2) \quad \frac{\gamma_{k+1}}{\gamma_k} > 1.$$

$$\text{Then } \frac{1}{k} - \kappa \frac{k+1}{k^2} \gamma_k > 0, \text{ i. e.}$$

$$\gamma_k < \frac{1}{\kappa} \frac{k}{k+1} < \frac{1}{\kappa}.$$

Further, from (3.11) we obtain that

$$\frac{\gamma_{k+1}}{\gamma_k} \leq \frac{k+1}{k} \leq 2$$

with $k \geq 1$. Now only two situations are possible.

(1) There is only a finite number of indices k for which $\gamma_k < \frac{1}{\kappa}$. Then due to the above statements for all great sequence $\{\gamma_k\}$ does not increase monotonically, i.e. remains bounded.

(2) There is an infinite number of indices k for which $\gamma_k < \frac{1}{\kappa}$.

We shall denote the set of such indices k by $\mathcal{Y}$ so that $\gamma_k \leq \frac{1}{\kappa}$ for $k \in \mathcal{Y}$. Let now $j \notin \mathcal{Y}$. Then there will be two indices $k_1, k_2 \in \mathcal{Y}$ such that $k_1 < j < k_2$ and $k \notin \mathcal{Y}$ for all $k_1 < k < k_2$. Then

$$\gamma_{k_1+1} \leq 2\gamma_{k_1} \leq \frac{2}{\kappa}$$

and $\gamma_{i+1} \leq \gamma_i$ for all $i = k_i + 1, \dots, k_2 - 1$. Therefore

$$\gamma_j \leq \frac{2}{\kappa}, \quad j \notin \mathcal{Y}.$$

This shows that in this case as well the sequence has a limit, a certain constant C . It follows that

$$\varphi_k = \frac{\gamma_k}{k} \leq \frac{C}{k}$$

and this completes the proof of the theorem.

The estimate obtained shows that the algorithm converges not very rapidly. But we have obtained the upper bound and it can seem that in reality the algorithm converges at a faster rate. Unfortunately, it is not so in general. As shown by M.D. Canon and C.D. Cullum, the bounds obtained are precise if the function being minimized is convex on a polyhedron.

Estimate of Convergence for a Strongly Convex Region

Let region Ω be strongly convex, i.e. there is a number $\delta > 0$ such that for any $x, y \in \Omega$ points $\frac{x+y}{2} + w$ belong to Ω for all the w such that $\|w\| \leq \delta \|x - y\|^2$. Then

$$\begin{aligned} \eta(x) = \min_{z \in \Omega} (f'(x), z - x) &\leq \min_{\|w\| \leq \delta \|z(x) - x\|^2} \left(f'(x), \frac{z(x) + x}{2} + w - x \right) \\ &\leq \frac{1}{2} (f'(x), z(x) - x) - \delta \|z(x) - x\|^2 \|f'(x)\|. \end{aligned}$$

Hence

$$\frac{1}{2} \eta(x) \leq -\delta \|f'(x)\| \|z(x) - x\|^2$$

or

$$-\frac{1}{2} \frac{\eta(x)}{\|p(x)\|^2} \geq \delta \|f'(x)\|. \quad (3.12)$$

Theorem 3.2. *If $f(x)$ is a convex function and region Ω is strongly convex and if $\|f'(x)\| \geq \varepsilon_0 > 0$ for all $x \in \Omega$, then the method of conditional gradient converges at the rate of a geometric progression, i.e. $\|x_k - x_*\| \leq Cq_0^k$, $q_0 < 1$.*

Proof. From (3.7) and (3.8) we obtain

$$\varphi_k - \varphi_{k+1} = f(x_k) - f(x_{k+1}) \geq \frac{1}{8L} \frac{\eta^2(x_k)}{\|p(x_k)\|^2}.$$

Using (3.12) and (3.10), we obtain

$$\varphi_k - \varphi_{k+1} \geq \frac{1}{4L} \delta \varepsilon_0 (-\eta(x)) \geq \frac{\delta \varepsilon_0}{4L} \varphi_k,$$

i.e.

$$\varphi_{k+1} \leq \left(1 - \frac{\delta \varepsilon_0}{4L}\right) \varphi_k.$$

Therefore

$$\varphi_m \leq q^m \varphi_0, \quad q = 1 - \frac{\delta \varepsilon_0}{4L} < 1.$$

Because of the necessary and sufficient conditions for a minimum, we have

$$(f'(x_*), x - x_*) \geq 0.$$

Therefore for all the w , $\|w\| \leq \delta \|x - x_*\|^2$,

$$\left(f'(x_*), \frac{x+x_*}{2} + w - x_*\right) \geq 0,$$

Hence

$$\frac{1}{2} (f'(x_*), x - x_*) \geq \delta \|x - x_*\|^2 \|f'(x_*)\|.$$

But $\|f'(x_*)\| \geq \varepsilon_0$ and since $f(x)$ is convex, we have

$$f(x) - f(x_*) \geq (f'(x_*), x - x_*).$$

Finally, we obtain

$$f(x) - f(x_*) \geq 2\delta\varepsilon_0 \|x - x_*\|^2. \quad (3.13)$$

Hence

$$\|x_k - x_*\| \leq \left(\frac{1}{2\delta\varepsilon_0}\right)^{1/2} \varphi_k^{1/2} \leq \left(\frac{\varphi_0}{2\delta\varepsilon_0}\right)^{1/2} (q^{1/2})^k.$$

With the notation

$$C = \left(\frac{\varphi_0}{2\delta\varepsilon_0}\right)^{1/2}, \quad q_0 = \left(1 - \frac{\delta\varepsilon_0}{4L}\right)^{1/2}$$

we obtain

$$\|x_k - x_*\| \leq C q_0^k.$$

Q.E.D.

Newton's Method with Step Adjustment

We consider now the problem of the minimization of a convex smooth function $f(x)$ in (a convex, compact) set Ω .

For solving this problem the iterative process

$$x_{k+1} = x_k + \alpha_k p_k, \quad \alpha_k > 0 \quad (3.14)$$

can be used in which the direction of motion $p_k = \bar{x}_k - x_k$ is the solution of the problem of minimization in set Ω of the quadratic function

$$\psi_k(x) = (f'(x_k), x - x_k) + \frac{1}{2} (f''(x_k)(x - x_k), x - x_k),$$

and as α_k we take the maximum value of parameter α , obtained by reducing the initial $\alpha = 1$ until the parameter satisfies the inequality

$$f(x_k + \alpha p_k) - f(x_k) \leq \varepsilon \alpha \psi_k(\bar{x}_k), \quad 0 < \varepsilon < 1. \quad (3.15)$$

Parameter α_k can be chosen by other methods analogous to those described in Sec. 2, Chap. II (methods (2.2), (2.3)).

We shall ascertain below that the rate of convergence of Newton's method under definite conditions is either superlinear, or quadratic. Consequently, if the problem of the minimization of function $\psi_k(x)$, $x \in \Omega$ can be solved easily enough, then Newton's method proves very effective.

Properties of Newton's Method

Theorem 3.3. *If for the minimization of a convex twice continuously differentiable function $f(x)$ in a convex closed bounded set Ω we use method (3.14) in which α_k and p_k are determined as described above, then (whatever the initial approximation $x_0 \in \Omega$ chosen):*

- (1) $f(x_k)$ decreases monotonically,
- (2) $\lim_{k \rightarrow \infty} f(x_k) = f(x_*) = \min_{x \in \Omega} f(x)$.

Proof. There is a minimum point $\bar{x}_k$ (possibly not unique) of the continuous function $\psi_k(x)$ in compact set Ω (Weierstrass' theorem). With any k point $x_{k+1} \in \Omega$ since $x_{k+1} = x_k + \alpha_k(\bar{x}_k - x_k) = \alpha_k \bar{x}_k + (1 - \alpha_k)x_k$ and $\alpha_k \in [0, 1]$. Since function $\psi_k(x)$ is convex, we have $\psi_k(x_{k+1}) = \psi_k(\alpha_k \bar{x}_k + (1 - \alpha_k)x_k) \leq \alpha_k \psi_k(\bar{x}_k) + (1 - \alpha_k)\psi_k(x_k)$. But $\psi_k(x_k) = 0$, therefore

$$\psi_k(x_{k+1}) \leq \alpha_k \psi_k(\bar{x}_k). \quad (3.16)$$

Now making use of Taylor's formula and of (3.16), we have:

$$\begin{aligned} f(x_{k+1}) - f(x_k) &= \psi_k(x_{k+1}) = \frac{\alpha_k^2}{2} (F_k p_k, p_k) \\ &\leq \alpha_k \psi_k(\bar{x}_k) \left(1 + \frac{\alpha_k}{2} \frac{\|F_k\| \|p_k\|^2}{\psi_k(\bar{x}_k)} \right) \end{aligned} \quad (3.17)$$

where $F_k = f''(x_{kc}) - f''(x_k)$, $x_{kc} = x_k + \theta(x_{k+1} - x_k)$, $\theta \in [0, 1]$. It follows that if $\psi_k(\bar{x}_k) \neq 0$ (in this case $\psi_k(\bar{x}_k) < \psi_k(x_k) = 0$), then with a certain $\alpha_k > 0$ the inequality

$$1 + \frac{\alpha_k}{2} \frac{\|F_k\| \|p_k\|^2}{\psi_k(\bar{x}_k)} \geq \varepsilon \quad (3.18)$$

holds. But at the same time the inequality (3.15) holds as well and this proves that the described method of choosing α_k may be applied.

It follows from (3.15) that $f(x_{k+1}) \leq f(x_k)$. We show now that $\psi_k(\bar{x}_k) \rightarrow 0$ as $k \rightarrow \infty$. In a closed bounded set Ω continuous function $f''(x)$ has an upper bound: $\|f''(x)\| \leq M$. Consequently, $\|F_k\| \leq 2M$, and vector p_k has an upper bound too:

$\|p_k\| \leq \max_{x, y \in \Omega} \|x - y\| = d$. Suppose that with any k we have $\psi_k(\bar{x}_k) \leq -\beta < 0$. Then

$$1 + \frac{\alpha_k}{2} \frac{\|F_k\| \|p_k\|^2}{\psi_k(\bar{x}_k)} \geq 1 - \alpha_k \frac{Md^2}{\beta},$$

and, hence, inequality (3.18) (and therefore (3.15) too) is always satisfied even with $\alpha_k = \frac{(1-\varepsilon)\beta}{Md^2} = C > 0$. But it follows from (3.15) that at the same time $f(x_{k+1}) - f(x_k) \leq -\varepsilon C\beta$ with any k and this contradicts the fact that in compact set Ω function $f(x)$ has a lower bound.

Thus the condition $\psi_k(\bar{x}_k) \leq -\beta$ with any k cannot be fulfilled, i.e. in any case as $k \rightarrow \infty$ the condition $\psi_k(\bar{x}_k) \rightarrow 0$ must be fulfilled. This means that at any limit point of sequence (3.14) the necessary (and sufficient, because of the convexity of $f(x)$) condition for a minimum of function $f(x)$ in set Ω (see Chap. I, Sec. 4) is fulfilled. Taking this into account, the last assumption of the theorem can be proved as in theorem 3.1.

The theorem just proved shows that in the problem under consideration, as distinct from unconstrained minimization problems when Newton's method can be applied only to minimize strongly convex functions, this method can be applied to minimize also convex functions, since set Ω is bounded. However, the application of Newton's method to the minimization of strongly convex functions is of the greatest interest, for just in this case the method converges to the solution at a fast rate.

Theorem 3.4. *If in addition to the conditions of theorem 3.3, function $f(x)$ is strongly convex, i.e.*

$$m \|y\|^2 \leq (f''(x)y, y) \leq M \|y\|^2, \quad m > 0, \quad x \in \Omega, \\ y \in E^n, \quad (3.19)$$

then sequence (3.14) converges to the solution at a superlinear rate (i.e. the estimate (2.5), Chap. II, holds).

Proof. The existence and uniqueness of the solution of the problem under consideration follow from the general results of Sec. 3, Chap. I. At point $\bar{x}_k$ the necessary condition for a minimum of function $\psi_k(x)$ in set Ω (Sec. 4 of Chap. I)

$$(\psi'_k(\bar{x}_k), \bar{x}_k - x_k) \leq 0$$

is fulfilled, i.e.

$$(f'(x_k), \bar{x}_k - x_k) + (f''(x_k)(\bar{x}_k - x_k), \bar{x}_k - x_k) \leq 0.$$

Hence

$$(f'(x_k), p_k) \leq - (f''(x_k) p_k, p_k). \quad (3.20)$$

Taking into account this and the left-hand estimate (3.19), we find that

$$\psi_k(\bar{x}_k) \leq -\frac{m}{2} \|p_k\|^2. \quad (3.21)$$

Using this estimate we obtain from (3.17)

$$f(x_{k+1}) - f(x_k) \leq \alpha_k \psi_k(\bar{x}_k) \left(1 - \frac{\alpha_k \|F_k\|}{m}\right). \quad (3.22)$$

Since $\psi_k(\bar{x}_k) \rightarrow 0$ (theorem 3.3), it follows from (3.21) that $\|p_k\| \rightarrow 0$ as $k \rightarrow \infty$. Hence, because of the uniform continuity of the second derivative $f''(x)$ on set Ω , we have that $\|F_k\| \rightarrow 0$. But (3.22) implies that from a certain $k = N_1(\varepsilon)$ on, inequality (3.15) will be satisfied with $\alpha_k \equiv 1$, i.e. method (3.14) is transformed into the usual Newton's method with a unity step. With $k > N_1(\varepsilon)$ taking into account the convexity of $\psi_k(x)$, we obtain

$$\begin{aligned} \psi_k(\bar{x}_k) &= \psi_k(x_{k+1}) \geq (f'(x_k), x_{k+1} - x_k) \\ &= (f'(x_{k-1}), x_{k+1} - x_k) + (f'(x_k) - f'(x_{k-1}), x_{k+1} - x_k). \end{aligned}$$

Transforming the last scalar product by means of Lagrange's formula for operators (Chap. I, Sec. 5) we obtain

$$\begin{aligned} \psi_k(\bar{x}_k) &= (f'(x_{k-1}) + f''(x_{k-1})(x_k - x_{k-1}), x_{k+1} - x_k) \\ &\quad + (\Phi_k(x_k - x_{k-1}), x_{k+1} - x_k) \end{aligned} \quad (3.23)$$

where

$$\Phi_k = f''(x_{k-1} + \theta_1(x_k - x_{k-1})) - f''(x_{k-1}), \quad \theta \in [0, 1].$$

Note that $(f'(x_{k-1}) + f''(x_{k-1})(x_k - x_{k-1}), x_{k+1} - x_k) = (\psi'_{k-1}(x_k), x_{k+1} - x_k)$. Since $\psi_{k-1}(x_k) = \min_{x \in \Omega} \psi_{k-1}(x)$, we shall have with any $x \in \Omega$ that $(\psi'_{k-1}(x_k), x - x_k) \geq 0$ (the necessary condition for a minimum). Consequently, $(\psi'_{k-1}(x_k), x_{k+1} - x_k) \geq 0$ holds and therefore it follows from (3.23) that

$$\begin{aligned} -\psi_k(\bar{x}_k) &\leq \|\Phi_k\| \|x_k - x_{k-1}\| \|x_k - x_{k+1}\| \\ &= \|\Phi_k\| \|p_{k-1}\| \|p_k\|. \end{aligned} \quad (3.24)$$

Comparing estimates (3.21) and (3.24), we obtain

$$\|x_{k+1} - x_k\| \leq \frac{2\|\Phi_k\|}{m} \|x_k - x_{k-1}\|. \quad (3.25)$$

Because of the uniform continuity of $f''(x)$ on set Ω , we have $\|\Phi_k\| \rightarrow 0$. Consequently, there is a number $N(\varepsilon)$ such that with $k \geq N(\varepsilon)$ we find $\lambda_k = \frac{2\|\Phi_k\|}{m} < 1$. Let us take $\|x_N - x_{N-1}\| = C_1$,

$1 - \lambda_N = \gamma > 0$. Then $\|x_i - x_{N+l}\| \leq \sum_{k=N+l}^{i-1} \|x_{k+1} - x_k\| \leq C_1 \lambda_N \lambda_{N+1} \dots$
 $\dots \lambda_{N+l} (1 + \lambda_{N+l+1} + \lambda_{N+l+1}^2 + \dots + \lambda_{N+l+1}^{i-(N+l+1)}) \leq \frac{C_1}{1 - \lambda_N} \lambda_N \dots$
 $\dots \lambda_{N+l} = C \lambda_N \dots \lambda_{N+l}$. Consequently, as $i, l \rightarrow \infty$, $\|x_i - x_{N+l}\| \rightarrow 0$,
 i.e. sequence $\{x_k\}$ is fundamental and, because of the completeness
 of space E^n , has a limit $x_* \in \Omega$, and

$$\|x_{N+l} - x_*\| \leq C \lambda_N \lambda_{N+1} \dots \lambda_{N+l}. \quad (3.26)$$

By theorem 3.3,

$$\lim_{k \rightarrow \infty} f(x_k) = f(x_*) = \min_{x \in \Omega} f(x).$$

Thus sequence (3.14) converges to the solution and the rate of convergence, as can be seen from estimate (3.26), is superlinear. The theorem is proved.

If the second derivative $f''(x)$ on set Ω satisfies Lipschitz' condition with constant R , then inequality (3.25) takes the form

$$\|x_{k+1} - x_k\| \leq \frac{2R}{m} \|x_k - x_{k-1}\|^2. \quad (3.27)$$

We shall use the notation $\beta_k = \frac{2R}{m} \|x_{k+1} - x_k\|$. Since $\|x_{k+1} - x_k\| \rightarrow 0$, then there is a number $L(\epsilon)$ such that with $k > L(\epsilon)$ we have $\beta_k < 1$. Taking into account (3.27) with $k > L$, we have

$$\beta_k \leq \beta_{k-1}^2 \leq \dots \leq \beta_L^{2^{k-L}}.$$

Consequently, for any $i > L + l$, $l = 0, 1, \dots$,

$$\|x_i - x_{L+l}\| \leq \sum_{k=L+l}^{i-1} \|x_{k+1} - x_k\| \leq \frac{m}{2R} \sum_{k=L+l}^{i-1} \beta_L^{2^{k-L}} = \frac{m}{2R} \sum_{s=l}^{i-1} \beta_L^{2^s}.$$

Since $x_k \rightarrow x_*$, we have $\|x_{L+l} - x_*\| = \lim_{i \rightarrow \infty} \|x_{L+l} - x_i\|$, i.e.

$$\|x_{L+l} - x_*\| \leq \frac{m}{2R} \sum_{s=l}^{\infty} \beta_L^{2^s}.$$

This estimate can be given the form

$$\|x_{L+l} - x_*\| \leq C \beta_L^{2^l}, \quad C < \infty$$

(taking into account that the series $\sum_{s=l}^{\infty} \beta_L^{2^s}$ converges). The estimate obtained means that the following theorem is valid.

Theorem 3.5. *If the conditions of theorem 3.4 are fulfilled and, besides, matrix $f''(x)$ on set Ω satisfies Lipschitz' condition with constant*

R , then sequence (3.14) (in which α_k and p_k are chosen by the method described above) converges to the solution at a quadratic rate.

We shall study now the properties of Newton's method with the choice of α_k under the condition of $f(x)$ having the minimum value in the direction of motion.

The argument we used to estimate the rate of convergence of Newton's method in unconstrained problems (Sec. 2, Chap. II) are not suited to this case (since the right-hand estimate (1.11) of Chap. II does not hold).

Lemma 3.3. *If function $f(x)$ for which conditions (3.19) are fulfilled is being minimized and α_k in method (3.14) is chosen under the condition*

$$f(x_k + \alpha_k p_k) = \min_{0 \leq \alpha \leq 1} f(x_k + \alpha p_k), \quad (3.28)$$

then $x_k \rightarrow x_*$ and $\alpha_k \rightarrow 1$ as $k \rightarrow \infty$.

Proof. By Taylor's formula

$$f(x_{k+1}) - f(x_k) = \alpha_k (f'(x_k), p_k) + \frac{\alpha_k^2}{2} (f''(x_k), p_k, p_k).$$

With the value of α_k satisfying (3.28), the right-hand side of the last equality considered as a function of the variable α must attain its minimum. Therefore, it can be easily ascertained, taking into account estimates (3.19) and (3.20), that

$$1 \geq \alpha_k \geq -\frac{(f'(x_k), p_k)}{M \|p_k\|^2} \geq \frac{m \|p_k\|^2}{M \|p_k\|^2} = \frac{m}{M}.$$

Thus $\alpha_k \geq C > 0$; therefore in the same way as in theorem 3.3 it can be shown that $\psi_k(\bar{x}_k) \rightarrow 0$, i.e. sequence (3.14) in which α_k is chosen under condition (3.28) converges to the solution. At the same time $\|p_k\| \rightarrow 0$ and $\|F_k\| \rightarrow 0$ (theorem 3.4).

We shall demonstrate that $\alpha_k \rightarrow 1$:

$$\begin{aligned} f(x_{k+1}) &= \psi_k(x_{k+1}) + \frac{\alpha_k^2}{2} (F_k p_k, p_k) \\ &= \psi_k(\bar{x}_k) + (\psi'_k(\bar{x}_k), x_{k+1} - \bar{x}_k) \\ &\quad + \frac{1}{2} (\psi''_k(\bar{x}_k) (x_{k+1} - \bar{x}_k), x_{k+1} - \bar{x}_k) + \frac{\alpha_k^2}{2} (F_k p_k, p_k). \end{aligned}$$

Taking into account that $x_{k+1} - \bar{x}_k = (\alpha_k - 1) p_k$ we obtain:

$$\begin{aligned} f(x_{k+1}) &= \psi_k(\bar{x}_k) + (\psi'_k(\bar{x}_k), x_{k+1} - \bar{x}_k) \\ &\quad + \frac{(\alpha_k - 1)^2}{2} (\psi''_k(\bar{x}_k) p_k, p_k) + \frac{\alpha_k^2}{2} (F_k p_k, p_k). \end{aligned}$$

Note $(\psi'_k(\bar{x}_k), x_{k+1} - \bar{x}_k) \geq 0$, $(\psi''_k(\bar{x}_k) p_k, p_k) = (f''(x_k) p_k, p_k) \geq m \|p_k\|^2$. At the same time $(F_k p_k, p_k) = o(\|p_k\|^2)$, (since $\|F_k\| \rightarrow 0$ as $\|p_k\| \rightarrow 0$). Consequently, the minimum of the difference $f(x_{k+1}) - \psi_k(\bar{x}_k)$ can be attained only as $\alpha_k \rightarrow 1$; otherwise, with any k we should have $1 - \alpha_k \geq \beta > 0$ and at the same time $f(x_{k+1}) - \psi_k(\bar{x}_k) = O(\|p_k\|^2) > 0$, whereas with $\alpha = 1$ the difference $f(x_k) - \psi_k(\bar{x}_k) = \frac{1}{2}(F_k p_k, p_k) = o(\|p_k\|^2)$, i.e.

with a sufficiently great k we should always have $f(\bar{x}_k) < f(x_{k+1})$; this contradicts the condition for choosing α_k . The lemma is proved.

Theorem 3.6. *If function $f(x)$ satisfies the requirements of theorem 3.4 and in method (3.14) parameter α_k is chosen under the condition (3.28), then $x_k \rightarrow x_*$ at a superlinear rate.*

Proof. By estimates (3.16) and (3.21),

$$\psi_k(x_{k+1}) \leq \alpha_k \psi_k(\bar{x}_k) \leq \alpha_k^2 \psi_k(\bar{x}_k) \leq -\frac{x_k^2}{2} m \|p_k\|^2,$$

i.e.

$$\psi_k(x_{k+1}) \leq -\frac{m}{2} \|x_{k+1} - x_k\|^2. \quad (3.29)$$

On the other hand, $\psi_k(x_{k+1}) \geq (f'(x_k), x_{k+1} - x_k) = (f'(x_k), x_{k+1} - \bar{x}_{k-1}) + (f'(x_k), \bar{x}_{k-1} - x_k)$. Since at point x_k the minimum of $f(x)$ is attained in the direction $\bar{x}_{k-1} - x_{k-1}$, we have $(f'(x_k), \bar{x}_{k-1} - x_k) \geq 0$. Hence, we have

$$\begin{aligned} \psi_k(x_{k+1}) &\geq (f'(x_k), x_{k+1} - \bar{x}_{k-1}) \\ &= (f'(x_{k-1}), x_{k+1} - \bar{x}_{k-1}) + (f'(x_k) - f'(x_{k-1}), x_{k+1} - \bar{x}_{k-1}). \end{aligned}$$

Expressing the last term on the right-hand side by means of Lagrange's formula for operators, we obtain after some transformation

$$\begin{aligned} \psi_k(x_{k+1}) &\geq (f'(x_{k-1}) + f''(x_{k-1})(x_k - x_{k-1}), x_{k+1} - \bar{x}_{k-1}) \\ &\quad + (\Phi_k(x_k - x_{k-1}), x_{k+1} - \bar{x}_{k-1}) \end{aligned}$$

where $\Phi_k = f''(x_{k-1} + \theta(x_k - x_{k-1})) - f''(x_{k-1})$, $\theta \in [0, 1]$. Taking into account that $x_k - x_{k-1} = \alpha_{k-1}(\bar{x}_{k-1} - x_{k-1})$, we find that

$$\begin{aligned} \psi_k(x_{k+1}) &\geq (f'(x_{k-1}) + f''(x_{k-1})(\bar{x}_{k-1} - x_{k-1}), x_{k+1} - \bar{x}_{k-1}) \\ &\quad + ((\alpha_{k-1} - 1) f''(x_{k-1})(\bar{x}_{k-1} - x_{k-1}) + \Phi_k(x_k - x_{k-1}), x_{k+1} - \bar{x}_{k-1}). \end{aligned}$$

Since

$$\psi_{k-1}(\bar{x}_{k-1}) = \min_{x \in \Omega} \psi_{k-1}(x),$$

we have $(\psi'_{k-1}(\bar{x}_{k-1}), \bar{x}_{k-1} - x) \leq 0$ for all $x \in \Omega$ (the necessary condition for a minimum). Consequently

$$\begin{aligned} (\psi'_{k-1}(\bar{x}_{k-1}), \bar{x}_{k-1} - x_{k+1}) \\ = (f'(x_{k-1}) + f''(x_{k-1})(\bar{x}_{k-1} - x_{k-1}), \bar{x}_{k-1} - x_{k+1}) \leq 0. \end{aligned}$$

Taking this into account we see that the following estimate holds:

$$\psi_k(x_{k+1}) \geq \left(\left(\frac{\alpha_{k-1} - 1}{\alpha_{k-1}} f''(x_{k-1}) + \Phi_k \right) (x_k - x_{k-1}), x_{k+1} - \bar{x}_{k-1} \right).$$

Hence, with the notation $\|[(\alpha_{k-1} - 1)/\alpha_{k-1}] f''(x_{k-1}) + \Phi_k\| = b_k$, we obtain

$$\begin{aligned} -\psi_k(x_{k+1}) &\leq b_k \|x_k - x_{k-1}\| \|x_{k+1} - \bar{x}_{k-1}\| \\ &\leq b_k \|x_k - x_{k-1}\| (\|x_{k+1} - x_k\| + \|\bar{x}_{k-1} - x_k\|). \end{aligned}$$

Taking now into account that

$$\bar{x}_{k-1} - x_k = \frac{1 - \alpha_{k-1}}{\alpha_{k-1}} (x_k - x_{k-1})$$

and denoting $[(1 - \alpha_{k-1})/\alpha_{k-1}] b_k = c_k$, we obtain

$$-\psi_k(x_{k+1}) \leq b_k \|x_k - x_{k-1}\| \|x_{k+1} - x_k\| + c_k \|x_k - x_{k-1}\|^2.$$

Since $\alpha_k \rightarrow 1$, $\|p_k\| \rightarrow 0$ (lemma 3.3), we have $b_k \rightarrow 0$, $c_k \rightarrow 0$. Comparing the estimate obtained with (3.29) we establish that

$$\|x_{k+1} - x_k\|^2 \leq \xi_k \|x_k - x_{k-1}\| \|x_{k+1} - x_k\| + \rho_k \|x_k - x_{k-1}\|^2$$

where

$$\xi_k = \frac{2b_k}{m}, \quad \rho_k = \frac{2c_k}{m}.$$

Finally, having solved the quadratic inequality obtained for $\|x_{k+1} - x_k\|$, we find that

$$\|x_{k+1} - x_k\| \leq \mu_k \|x_k - x_{k-1}\|$$

where

$$\mu_k = \frac{\xi_k}{2} + \sqrt{\frac{\xi_k^2}{4} + \rho_k} \rightarrow 0 \text{ as } k \rightarrow \infty.$$

The remaining part of the proof is performed just as in theorem 3.4.

4. CUTTING HYPERPLANE METHOD

The cutting hyperplane method is meant for solving problems of convex programming. The basic idea of the method is that the admissible domain is approximated by a certain polyhedron which

diminishes from one iteration step to the next giving a better and better approximation to the admissible region about the solution.

The method is applied to the problem of convex programming formulated as follows: to minimize $f_0(x) = (c, x)$ with the constraint

$$f(x) \leq 0 \quad (4.1)$$

where $f(x)$ is a continuous convex function.

The fact that function $f_0(x)$ to be minimized is linear and the constraint (4.1) consists of only one inequality is of no great consequence, since if the region is defined by several inequalities

$$f_i(x) \leq 0, \quad i = 1, \dots, m \quad (4.2)$$

with convex functions $f_i(x)$, then the system of inequalities can be rewritten in the form (4.1) setting

$$f(x) = \max_{1 \leq i \leq m} f_i(x).$$

If convex function $f_0(x)$ is nonlinear, then by introducing an additional coordinate x^{n+1} and adding the inequality

$$f_{m+1}(x, x^{n+1}) = f_0(x) - x^{n+1} \leq 0$$

to the system (4.2), we can reduce the problem to the minimization of the linear function of x^{n+1} with constraints (4.2). Therefore, we shall study the problem in the form (4.1). Before going on to the description of the algorithm we remind the reader that vector a is a support vector to $f(x)$ at point x_0 if $f(x) \geq f(x_0) + (a, x - x_0)$ for all x . It follows from the results of Sec. 2, Chap. I that for a continuous convex function the set of such vectors is nonempty at any point of space.

Algorithm

Let

$$\Omega = \{x: f(x) \leq 0\}$$

be a nonempty admissible region. Suppose also that Ω is compact and vectors such as a_k , $k = -l, -(l-1), \dots, -1, 0$ and numbers b_k are known and that region

$$S = \{x: (a_k, x) - b_k \leq 0, \quad k = -l, \dots, 0\}$$

is compact and contains Ω .

For $k \geq 0$ the successive approximations are determined by the following rule. We set $S_0 = S$. If S_k has been constructed, then x_k is any solution of the problem of linear programming: to minimize $f_0(x) = (c, x)$ with $x \in S_k$. The next region S_{k+1} is constructed by

the following rule:

$$S_{k+1} = \{x: (a_{k+1}, x) - b_{k+1} \leq 0\} \cap S_k \quad (4.3)$$

where a_{k+1} is a support vector to $f(x)$ at point x_k and

$$b_{k+1} = (a_{k+1}, x_k) - f(x_k). \quad (4.4)$$

It follows from (4.3) that $S_{k+1} \subset S_k$ and for $k \geq 1$

$$S_k = \{x: (a_j, x) - b_j \leq 0, \\ j = -l, \dots, -1, 0, \dots, k-1\}. \quad (4.5)$$

Lemma 4.1. *For all $k \geq 1$, $\Omega \subset S_k$.*

Proof. Let $x \in \Omega$, i.e. $f(x) \leq 0$. Then

$$f(x) \geq f(x_{j-1}) + (a_j, x - x_{j-1}) = (a_j, x) - b_j$$

and, consequently, $(a_j, x) - b_j \leq 0$, $j = 1, \dots, k$. With $j \leq 0$ the last inequalities hold by virtue of choosing a_j and b_j for $j \leq 0$. The lemma is proved.

It follows directly from lemma 4.1 that

$$f_0(x_0) \leq f_0(x_1) \leq \dots \leq f_0(x_k) \leq f_0(x_{k+1}) \leq \dots$$

On the other hand, if x_* is the minimum point of $f_0(x)$ in Ω , then $f_0(x_k) \leq f_0(x_*)$ since $S_k \supset \Omega$.

Theorem 4.1. *Let $f(x)$ be a continuous convex function, region Ω be compact and there be a number K such that with each $x \in S$ vector a which is a support vector to $f(x)$ at point x satisfies the inequality $\|a\| \leq K$. Then any limit point x_* of sequence $\{x_k\}$, $k = 0, 1, \dots$, is a solution of problem (4.1) and $f(x_k) \rightarrow 0$.*

Proof. Since $S_0 = S$, $S_k \supset S_{k+1}$, the whole sequence $\{x_k\}$ belongs to the compact set S . Therefore there are always limit points of this sequence.

Note now that if $f(x_k) \leq 0$ for a certain k , then $x_k \in \Omega$ and, consequently, $f_0(x_k) \geq f_0(x_*)$. However, as was shown, $f_0(x_k) \leq f_0(x_*)$. Thus, $f_0(x_k) = f_0(x_*)$, i.e. x_k is the solution of the original problem.

Let now sequence $\{x_k\}$ be infinite and $f(x_k) > 0$ for all k . We shall show that $f(x_k) \rightarrow 0$. Suppose that the contrary is true. Then there is a number $r > 0$ and a subsequence of indices k (we denote it by $\mathcal{J}$) such that $f(x_k) \geq r$, $k \in \mathcal{J}$. Without loss of generality, we can take that $x_k \rightarrow \bar{x}$, $k \in \mathcal{J}$ since $\{x_k\}$ belongs to a compact set.

Let now k and j belong to $\mathcal{J}$ and $k > j$. Then by construction, point x_k satisfies the inequality

$$(a_{j+1}, x_k) - b_j = (a_{j+1}, x_k - x_j) + f(x_j) \leq 0;$$

hence

$$f(x_j) \leq (a_{j+1}, x_j - x_k) \leq K \|x_j - x_k\|.$$

But $\{x_k\}$, $k \in \mathcal{J}$ converges to $\bar{x}$ and therefore $\|x_j - x_k\| \leq r/(2K)$ for all sufficiently great k and j and so $f(x_j) \leq r/2$ for great j , and this contradicts the fact that $f(x_j) \geq r$, $j \in \mathcal{J}$.

Thus we have shown that $f(x_k) \rightarrow 0$. Let now $\bar{x}$ be any limit point, i.e. $x_k \rightarrow \bar{x}$, $k \in \mathcal{J}$, where $\mathcal{J}$ is a subsequence of indices. Then because of the continuity of $f(x)$,

$$f(\bar{x}) = \lim_{k \in \mathcal{J}} f(x_k) = 0,$$

i.e. $\bar{x} \in \Omega$. On the other hand, $f_0(x_k) \leq f_0(x_*)$ and therefore $f_0(\bar{x}) \leq f_0(x_*)$; it follows directly that $f_0(x) = f_0(\bar{x}_*)$ and $\bar{x}$ is also a solution of problem (4.1). The theorem is proved.

Computational Aspects

The algorithm of the cutting hyperplane method requires at each step the solution of the problem of linear programming: to minimize $f_0(x) = (c, x)$ with constraints

$$(a_i, x) - b_i \leq 0, \quad i = -l, \dots, k. \quad (4.6)$$

Thus the size of the problem being solved increases at every step. The computer memory required for storing vectors a_k also increases. In order to simplify the solving of problem (4.6), it is expedient to solve instead the dual problem which in this case takes the following

form: to maximize $-\sum_{i=-l}^k u^i b_i$ with constraints

$$\sum_{i=-l}^k u^i a_i + c = 0, \quad u^i \geq 0, \quad i = -l, \dots, k.$$

In solving this problem by the simplex-method the solution of the preceding problem serves as the trial solution for the next one. It is also a good approximation so that the solution of the new problem is obtained after a small number of iterations.

At each step of the algorithm it is necessary to compute vector a_{k+1} , support vector to $f(x)$ at point x_k . Recall (see Chap. I) that if $f(x)$ is a differentiable function, then $a_{k+1} = f'(x_k)$. If $f(x)$ is the maximum of the functions being differentiated, i.e. $f(x) = \max_{1 \leq i \leq m} f_i(x)$, then we can take as a_{k+1} any vector of the form

$\sum_{i \in \mathcal{J}(x_k)} \lambda_i f'_i(x_k)$, where $\lambda_i \geq 0$, $\sum_{i \in \mathcal{J}(x_k)} \lambda_i = 1$, $\mathcal{J}(x_k) = \{i: f_i(x_k) = f(x_k), 1 \leq i \leq m\}$; in particular, we can take $a_{k+1} = f'_i(x_k)$, where i is any index from $\mathcal{J}(x_k)$.

The foregoing follows from what was set forth in Sec. 2, Chap. I.

Concluding Remarks

In describing the cutting hyperplane method we followed J. E. Kelly's paper. At present there are many modifications of this method. They can be found in a paper by E. S. Levitin and B. T. Polyak. However, all these modifications do not seem to enhance the main property which is of interest to us, viz. the rate of convergence which has not been precisely estimated for the method described; however, the results obtained in the paper mentioned permit to form the judgement that this rate is not even that of a geometric progression.

5. LINEARIZATION METHOD

In this section we shall describe the method of solving the general problem of mathematical programming without making any assumption concerning the convexity of the functions to be dealt with. An important property of this method is the possibility of taking into account nonlinear equality constraints, this being a stumbling-block for most other methods.

It is required to minimize function $f_0(x)$, $x \in E^n$ with constraints

$$f_i(x) \leq 0, \quad i \in \mathcal{J}^-, \quad f_i(x) = 0, \quad i \in \mathcal{J}^0 \quad (5.1)$$

where $\mathcal{J}^-$ and $\mathcal{J}^0$ are finite sets of indices. We assume that all the functions $f_i(x)$ are continuously differentiable. (More fully the constraints with which the problem is considered will be specified below.) At point x_0 we substitute linear constraints for all (5.1) and a linear function for $f_0(x)$ by linearizing $f_i(x)$ at point x . As a result we obtain a problem of linear programming. It would be natural to take the solution of the linearized problem as the next approximation as we do in Newton's method for solving systems of nonlinear equations. Unfortunately, this way does not lead directly to the aim since as a rule the subsidiary problem of linear programming has no solution. Therefore, it is necessary to impose certain constraints on the increase of vector x at x_0 in order that the solution of the linearized problem should not shift too far from x_0 and should remain in the neighbourhood of x_0 such that linearization would still hold at it. This will be performed below by adding a quadratic term to the linearized objective function.

Note that each of the equalities $f_i(x) = 0$ is equivalent to the following two inequalities

$$f_i(x) \leq 0, \quad -f_i(x) \leq 0.$$

Therefore we can limit ourselves to considering only the case with inequality constraints. Such a constraint is convenient at least

in the theoretical substantiation of the algorithm though the doubling of the number of inequalities can be inconvenient in calculations. We shall give below the theoretical substantiation of the algorithm for the problem of the minimization of $f_0(x)$ with constraints

$$f_i(x) \leq 0, \quad i \in \mathcal{J}. \quad (5.2)$$

A modification of the algorithm for the general problem (5.1) will be considered separately.

Thus we shall study the algorithm for problem (5.2) without loss of generality. Clearly, we can always assume that among the inequalities (5.2), there is the trivial one: $0 \leq 0$. Therefore it will be assumed that among the functions $f_i(x)$, $i \in \mathcal{J}$ there is one identically equal to zero: $f_i(x) \equiv 0$.

Basic Assumptions

We set

$$F(x) = \max_{i \in \mathcal{J}} f_i(x)$$

$$\mathcal{J}_\delta(x) = \{i \in \mathcal{J} : f_i(x) \geq F(x) - \delta\}, \quad \delta \geq 0. \quad (5.3)$$

By assumption, $F(x) \geq 0$ with all the x . Suppose that there are constants $N > 0$, $\delta > 0$ such that:

(a) the set

$$\Omega_N = \{x : f_0(x) + NF(x) \leq C_0\}, \quad C_0 = f_0(x_0) + NF(x_0)$$

is bounded;

(b) the gradients of functions $f_i(x)$, $i \in \{0\} \cup \mathcal{J}$ in Ω_N satisfy Lipschitz' condition, i.e.

$$\|f'_i(x_1) - f'_i(x_2)\| \leq L \|x_1 - x_2\|,$$

(c) the problem of quadratic programming

$$\min (f'_0(x), p) + \frac{1}{2} \|p\|^2,$$

$$(f'_i(x), p) + f_i(x) \leq 0, \quad i \in \mathcal{J}_\delta(x) \quad (5.4)$$

is solvable for $p \in E^n$ with any $x \in \Omega_N$ and there are Lagrange multipliers $u^i(x)$, $i \in \mathcal{J}_\delta(x)$ such that $\sum_{i \in \mathcal{J}_\delta(x)} u^i(x) \leq N$. In this section,

$\|p\|$ will always denote the Euclidean norm of vector p .

In what follows we shall denote the solution of problem (5.4) by $p(x)$ and Lagrange multipliers by $u^i(x)$, $i \in \mathcal{J}_\delta(x)$.

Formulation of the Algorithm

Let x_0 be the initial approximation and we take ε such that $0 < \varepsilon < 1$. Let point x_k be already constructed by the algorithm. The construction of the next approximation will be performed in two stages:

(1) We solve problem (5.4) with $x = x_k$ and find its solution, vector $p_k \equiv p(x_k)$.

(2) We find the first of the values of $i = 0, 1, \dots$, satisfying the inequality

$$f\left(x_k + \frac{1}{2^i} p_k\right) + NF\left(x_k + \frac{1}{2^i} p_k\right) \leq f(x_k) + NF(x_k) - \frac{1}{2^i} \varepsilon \|p_k\|^2.$$

If this inequality is satisfied for the first time with $i = i_0$, then we take $\alpha_k = 2^{-i_0}$, $x_{k+1} = x_k + \alpha_k p_k$.

Thus the following inequality is satisfied at each step:

$$f(x_{k+1}) + NF(x_{k+1}) \leq f(x_k) + NF(x_k) - \alpha_k \varepsilon \|p_k\|^2. \quad (5.5)$$

Convergence of the Algorithm

We show that the choice of the step α_k at each iteration is performed after a finite number of successive halvings of unity and substantiate the convergence of the algorithm.

From the results of Sec. 3, Chap. I it follows that $p(x)$ is the solution of problem (5.4) if and only if there are $u^i(x) \geq 0$, $i \in \mathcal{J}_\delta(x)$ such that

$$\begin{aligned} f'_0(x) + p(x) + \sum_{i \in \mathcal{J}_\delta(x)} u^i(x) f'_i(x) &= 0, \\ u^i(x) ((f'_i(x), p(x)) + f_i(x)) &= 0, \quad i \in \mathcal{J}_\delta(x). \end{aligned} \quad (5.6)$$

Therefore

$$\begin{aligned} (f'_0(x), p(x)) &= - \sum_{i \in \mathcal{J}_\delta(x)} u^i(x) (f'_i(x), p(x)) - \|p(x)\|^2 \\ &= \sum_{i \in \mathcal{J}_\delta(x)} u^i(x) f_i(x) - \|p(x)\|^2. \end{aligned} \quad (5.7)$$

Lemma 5.1. *In order that point x satisfy inequalities (5.2) and the necessary conditions for the minimum of $f_0(x)$ with constraints (5.2), it is necessary and sufficient that the equality $p(x) = 0$ be satisfied.*

Proof. Let point x satisfy (5.2) and the necessary conditions for the minimum of $f_0(x)$. Then there are numbers $u^i \geq 0$, $i \in \mathcal{J}$ such that

$$f'_0(x) + \sum_{i \in \mathcal{J}} u^i f'_i(x) = 0, \quad u^i f_i(x) = 0, \quad i \in \mathcal{J}. \quad (5.8)$$

If x satisfies (5.2), then $F(x) = 0$ and therefore $\mathcal{J}_0(x)$ coincides with the set of i for which $f_i(x) = 0$. Besides, from the second of relations (5.8), $u^i = 0$ if $f_i(x) < 0$, i.e. if $i \notin \mathcal{J}_0(x)$. Therefore taking into account that $\mathcal{J}_\delta(x) \supseteq \mathcal{J}_0(x)$ we can rewrite (5.8) in the following form:

$$f'_0(x) + \sum_{i \in \mathcal{J}_\delta(x)} u^i f'_i(x) = 0, \quad u^i f_i(x) = 0, \quad i \in \mathcal{J}_\delta(x).$$

The comparison of the last expressions with (5.6) shows that vector $p = 0$ is the solution of problem (5.4) for with $p = 0$ all constraints (5.4) are satisfied (since (5.2) are satisfied) and the fulfilment of (5.6) with $p = 0$ is the necessary and sufficient condition for vector $p = 0$ to be the solution of (5.4).

Let now $p(x) = 0$. This means that the constraints of (5.4) are satisfied with $p = 0$, i.e. $f_i(x) \leq 0$, $i \in \mathcal{J}_\delta(x)$. Since for $i \notin \mathcal{J}_\delta(x)$ we have

$$f_i(x) \leq F(x) - \delta \leq f_j(x) \leq 0$$

where $j \in \mathcal{J}_\delta(x)$, then point x satisfies all constraints (5.2). Besides with $p = 0$ relations (5.6) are transformed into (5.8) if we take $u^i = 0$, $i \notin \mathcal{J}_\delta(x)$. Thus, the necessary conditions for the minimum of $f_0(x)$ with constraints (5.2) are satisfied too and this completes the proof.

Let us now estimate the changes of all the functions comprised by the problem with a shift from point x_k in direction p_k .

For $i \in \mathcal{J}_\delta(x_k)$ using Taylor's formula we obtain:

$$f_i(x_k + \alpha p_k) = f_i(x_k) + \alpha(p_k, f'_i(x_k)) + \alpha(p_k, f'_i(\theta_i) - f'_i(x_k))$$

where $\theta_i = x_k + \alpha \xi_i p_k$, $0 \leq \xi_i \leq 1$. Since p_k is the solution of (5.4) with $x = x_k$, we have

$$\begin{aligned} f_i(x_k + \alpha p_k) &\leq f_i(x_k) - \alpha f_i(x_k) + \alpha^2 \|p_k\|^2 L \\ &\leq (1 - \alpha) f_i(x_k) + \alpha^2 \|p_k\|^2 L \end{aligned} \quad (5.9)$$

in deriving which we made use of the fact that the gradients of $f_i(x)$ satisfy Lipschitz' condition.

For $i \notin \mathcal{J}_\delta(x_k)$,

$$\begin{aligned} f_i(x_k + \alpha p_k) &= f_i(x_k) + \alpha(p_k, f'_i(\theta_i)) \\ &\leq F(x_k) - \delta + \alpha K \|p_k\| \end{aligned} \quad (5.10)$$

where K is a quantity which bounds $\|f'_i(x)\|$ in Ω_N .

Since

$$(1 - \alpha) F(x_k) \geq F(x_k) - \delta + \alpha K \|p_k\|$$

for α such that $\alpha \leq 1$,

$$0 \leq \alpha \leq \frac{\delta}{F(x_k) + K \|p_k\|}, \quad (5.11)$$

it follows from (5.9) and (5.10) that for all the i the following inequality holds:

$$f_i(x_k + \alpha p_k) \leq (1 - \alpha) F(x_k) + \alpha^2 L \|p_k\|^2 \quad (5.12)$$

provided α satisfies condition (5.11).

By analogy to the preceding estimates

$$\begin{aligned} f_0(x_k + \alpha p_k) &= f_0(x_k) + \alpha(p_k, f'_0(x_k)) + \alpha(p_k, f'_0(\theta_0) - f'_0(x_k)), \\ \theta_0 &= x_k + \alpha \xi_0 p_k, \quad 0 \leq \xi_0 \leq 1. \end{aligned}$$

Using (5.7) and Lipschitz' condition for gradients we obtain

$$f_0(x_k + \alpha p_k) \leq f_0(x_k) + \alpha \left(\sum_{i \in \mathcal{J}_\delta(x_k)} u^i(x_k) f_i(x_k) \right) - \alpha \|p_k\|^2 + \alpha^2 L \|p_k\|^2.$$

Hence and from (5.12) it follows that

$$\begin{aligned} f_0(x_k + \alpha p_k) + NF(x_k + \alpha p_k) &\leq f_0(x_k) + NF(x_k) \\ &\quad + \alpha \left(\sum_{i \in \mathcal{J}_\delta(x_k)} u^i(x_k) f_i(x_k) - NF(x_k) \right) - \alpha \|p_k\|^2 \\ &\quad + \alpha^2 (N + 1) L \|p_k\|^2. \end{aligned} \quad (5.13)$$

Recall now that $u^i(x_k) \geq 0$, $F(x_k) \geq 0$ and

$$\sum_{i \in \mathcal{J}_\delta(x_k)} u^i(x_k) \leq N.$$

Therefore

$$\sum_{i \in \mathcal{J}_\delta(x_k)} u^i(x_k) f_i(x_k) - NF(x_k) \leq 0.$$

But then (5.13) can be written in the form

$$\begin{aligned} f_0(x_k + \alpha p_k) + NF(x_k + \alpha p_k) &\leq f_0(x_k) + NF(x_k) \\ &\quad - \alpha \|p_k\|^2 (1 - \alpha (N + 1) L) \end{aligned}$$

or, if

$$0 \leq \alpha \leq \frac{1 - \varepsilon}{(N + 1) L} \quad (5.14)$$

then

$$\begin{aligned} f_0(x_k + \alpha p_k) + NF(x_k + \alpha p_k) &\leq f_0(x_k) + NF(x_k) \\ &\quad - \alpha \varepsilon \|p_k\|^2. \end{aligned} \quad (5.15)$$

Thus if

$$\begin{aligned} 0 &\leq \alpha \leq \bar{\alpha}_k, \\ \bar{\alpha}_k &= \min \left(1, \frac{\delta}{F(x_k) + K \|p_k\|}, \frac{1 - \varepsilon}{(N + 1) L} \right), \end{aligned}$$

then the inequality (5.15) holds.

But this means that inequality (5.5) is satisfied after a finite number of trials of $\alpha = 2^{-i}$, $i = 0, 1, \dots$, and with the inequality

$$\alpha_k > \frac{1}{2} \bar{\alpha}_k. \quad (5.16)$$

We shall prove now the following theorem on the convergence of the process.

Theorem 5.1. *If the assumptions of the subsection on p. 189 hold, then the process has the following properties:*

- (a) $F(x_k) \rightarrow 0$ as $k \rightarrow \infty$;
- (b) at any limit point x_* of sequence $\{x_k\}$, $k = 0, 1, \dots$, inequalities (5.2) are satisfied and also the conditions for the minimum of $f_0(x)$ with constraints (5.2).

Remark. The fact that $F(x_k)$ tends to zero means that sequence $\{x_k\}$ satisfies constraints (5.2) more and more precisely.

Proof. All points of $\{x_k\}$ belong to Ω_N since function $f_0(x) + NF(x)$, by (5.15), decreases from step to step. Further, as Ω_N is a compact set, $f_0(x) + NF(x)$ is bounded in this set since the function is continuous. Hence

$$\alpha_k \|p_k\|^2 \rightarrow 0 \quad (5.17)$$

as $k \rightarrow \infty$ for otherwise $f_0(x) + NF(x)$ decreases without limit along sequence $\{x_k\}$.

We shall prove now that $p_k \rightarrow 0$. Indeed, if p_k does not tend to zero, then it follows from (5.17) that $\alpha_k \rightarrow 0$ along a certain subsequence of indices k . But it follows from (5.16) and the expression of $\bar{\alpha}_k$ that then for great k

$$\alpha_k \geq \frac{1}{2} \bar{\alpha}_k = \frac{1}{2} \frac{\delta}{F(x_k) + K \|p_k\|}.$$

Therefore, the right-hand part of the last inequality must tend to zero. As $F(x)$ is a continuous function in the compact set Ω_N , $F(x)$ has an upper bound and the expression $\frac{\delta}{F(x_k) + K \|p_k\|}$ can tend to zero only if $\|p_k\| \rightarrow +\infty$. But from (5.6) we obtain that

$$\|p(x_k)\| = \left\| f'_0(x_k) + \sum_{i \in \mathcal{J}_\delta(x_k)} u^i(x_k) f'_i(x_k) \right\| \leq K(N+1).$$

Thus we have come to a contradiction through the assumption that p_k does not tend to zero.

By definition of p_k , the relations

$$(f'_i(x_k), p_k) + f_i(x_k) \leq 0, \quad i \in \mathcal{J}_\delta(x_k)$$

hold. Therefore

$$f_i(x_k) \leq -(f'_i(x_k), p_k) \leq K \|p_k\|, \quad i \in \mathcal{J}_\delta(x_k).$$

But $f_j(x_k) \leq f_i(x_k)$, $j \in \mathcal{J}_\delta(x_k)$, $i \in \mathcal{J}_\delta(x_k)$. Hence

$$F(x_k) = \max_{i \in \mathcal{J}} f_i(x_k) \leq K \|p_k\|.$$

Consequently, $F(x_k) \rightarrow 0$ as $k \rightarrow \infty$ for $F(x_k) \geq 0$. Further, let us take $u^i(x) = 0$, $i \in \mathcal{J}_\delta(x)$. Then we can rewrite (5.6) along sequence $\{x_k\}$ in the following form:

$$\begin{aligned} f'_0(x_k) + p_k + \sum_{i \in \mathcal{J}} u^i(x_k) f'_i(x_k) &= 0, \\ u^i(x_k) ((f'_i(x_k), p_k) + f_i(x_k)) &= 0, \quad i \in \mathcal{J}. \end{aligned} \quad (5.18)$$

Let now x_* be a limit point of $\{x_k\}$. As $x_k \in \Omega_N$ and Ω_N is compact, there are always such points. Without loss of generality, we may take that $x_k \rightarrow x_*$. Besides, since $u^i(x) \geq 0$, $i \in \mathcal{J}$ and their sum is limited, we can take that $u^i(x_k) \rightarrow u^i$ as $k \rightarrow \infty$.

Taking the limit in (5.18) we obtain:

$$f'_0(x_*) + \sum_{i \in \mathcal{J}} u^i f'_i(x_*) = 0, \quad u^i f_i(x_*) = 0, \quad i \in \mathcal{J}.$$

Besides, $u^i \geq 0$ since $u^i(x_k) \geq 0$ and point x_* satisfies all the constraints (5.2) since $f_i(x_k) \leq F(x_k)$ and $F(x_k) \rightarrow 0$. Hence taking the limit, we obtain $f_i(x_*) \leq 0$. Thus we have ascertained that the necessary conditions for a minimum are fulfilled at point x_* . The theorem is proved.

Corollary. If the only point at which the necessary conditions for a minimum are fulfilled is the minimum point, then the sequence generated by the algorithm converges to the minimum point of $f_0(x)$ with constraints (5.2).

Indeed, in this case by theorem 5.1, the only limit point of sequence $\{x_k\}$ can be but the minimum point.

Computational Aspects

The basic operation which requires considerable computations at every step in implementing the algorithm is the solving of problem (5.4). This is a problem of quadratic programming. In choosing the method for solving this problem it is necessary to take into account that problem (5.4) must be solved after a finite number of steps, since it is a subsidiary problem. Besides, since constant N is not known beforehand, it is expedient to obtain the corresponding Lagrange multipliers $u^i(x)$ in solving problem (5.4) in order to check whether the choice of N was right or not. On these grounds it seems expedient to pass to the dual problem and to solve it by the method of conjugate gradients which was discussed in the subsection on p. 160.

We shall construct now the dual of problem (5.4). As stated in Chap. I, Sec. 3, the objective function of the dual problem has the following form:

$$\varphi(u) = \min_p \left[(f'_0(x), p) + \frac{1}{2} \|p\|^2 + \sum_{i \in \mathcal{J}_\delta(x)} u^i ((f'_i(x), p) + f_i(x)) \right]. \quad (5.19)$$

Equating to zero the derivatives with respect to p of the right-hand side of the last equality, we find that the minimum is attained with

$$p = -f'_0(x) - \sum_{i \in \mathcal{J}_\delta(x)} u^i f'_i(x). \quad (5.20)$$

Thus, point p is uniquely defined by vector u with components u^i , $i \in \mathcal{J}_\delta(x)$.

Substituting (5.20) into the right-hand side of (5.19), we obtain

$$\varphi(u) = -\frac{1}{2} \left\| f'_0(x) + \sum_{i \in \mathcal{J}_\delta(x)} u^i f'_i(x) \right\|^2 + \sum_{i \in \mathcal{J}_\delta(x)} u^i f_i(x). \quad (5.21)$$

Thus, we have calculated the objective function of the dual problem. The dual problem consists now in the maximization of $\varphi(u)$ with constraints $u^i \geq 0$, $i \in \mathcal{J}_\delta(x)$.

Thus we have come to a problem of the maximization of a quadratic form with simple constraints; it is expedient to solve this problem by the method of conjugate gradients (the subsection on p. 160).

As a result of the solution of the dual problem we obtain Lagrange multipliers $u^i(x)$ and according to what was stated in Sec. 3, Chap. I, the substitution of $u^i(x)$ into (5.20) gives vector $p(x)$, the solution of the primal problem.

Another problem is the choosing of constants N and δ . Speaking generally, the quantity N is unknown. Choosing it too great can be a disadvantage, since by formula (5.14), this can involve a considerable reduction of the step. Therefore, it is expedient to estimate N during the implementation of the algorithm. For example, if at a certain step it occurred that

$$N \leq \sum_{i \in \mathcal{J}_\delta(x_k)} u^i(x_k),$$

then N should be changed to

$$N = 2 \sum_{i \in \mathcal{J}_\delta(x_k)} u^i(x_k). \quad (5.22)$$

Experience shows that such a correction brings success. Besides, it is clear from theoretical reasons that if x_k is sufficiently close to the

limit point, then in the regular case $u^i(x_k)$ proves close to Lagrange multipliers at point x_* which is the solution of the problem and therefore formula (5.22) leads to success. The behaviour of factors $u^i(x_k)$ will be considered at greater length below.

As to quantity δ , it should be reduced if subsidiary problem (5.4) proves insolvable at a certain step.

We shall now describe the conditions under which there exist constants N and δ . In fact, they exist in a considerably broader class of problems.

Theorem 5.2. *Let all functions $f_0(x)$, $f_i(x)$, $i \in \mathcal{J}$ be convex and there be a point $\bar{x}$ such that*

$$f_i(\bar{x}) < 0, \quad i \in \mathcal{J}.$$

Besides, let $f_0(x)$ tend to $+\infty$ as $x \rightarrow +\infty$ and let point x_0 satisfy constraints (5.2). Then with any $\delta > 0$, multipliers $u^i(x)$, $i \in \mathcal{J}_\delta(x)$ have an upper bound on set Ω_N with sufficiently great N and if Ω_N is compact.

Proof. Recall that

$$\Omega_N = \{x: f_0(x) + NF(x) \leq f_0(x_0) + NF(x_0)\}.$$

Hence and from the continuity of $f_0(x)$ and $F(x)$, it follows that Ω_N is a closed set. On the other hand, Ω_N is bounded since by hypothesis, $f_0(x) \rightarrow +\infty$ as $x \rightarrow +\infty$ and therefore

$$f_0(x) + NF(x) > f_0(x_0) + NF(x_0)$$

with all x sufficiently great in norm. Further, since x_0 satisfies (5.2), then $F(x_0) = 0$. Therefore with all N we have $\Omega_N \subset \Omega_0$. Indeed, it follows from $x \in \Omega_N$ that

$$f_0(x) \leq f_0(x) + NF(x) \leq f_0(x_0) + NF(x_0) = f_0(x_0),$$

i.e. $f_0(x) \leq f_0(x_0)$. Obviously, set Ω_0 is also compact by the assumptions of the theorem.

Further, since all $f_i(x)$ are convex we have

$$f_i(x) + (f'_i(x), \bar{x} - x) \leq f_i(\bar{x}) < 0. \quad (5.23)$$

Therefore, the system of constraints of problem (5.4) is consistent with any $\delta > 0$, as vector $\bar{p} = \bar{x} - x$ satisfies it.

Let now $u^i(x)$ be Lagrange multipliers of problem (5.4). Then by Kuhn-Tucker's theorem

$$\begin{aligned} \frac{1}{2} \|p(x)\|^2 + (f'_0(x), p(x)) &\leq \frac{1}{2} \|p\|^2 + (f'_0(x), p) \\ &+ \sum_{i \in \mathcal{J}_\delta(x)} u^i(x) ((f'_i(x), p) + f_i(x)) \end{aligned}$$

for all p . In particular, with $p = \bar{p} \equiv \bar{x} - x$, by (5.23), we have

$$\begin{aligned} & \frac{1}{2} \|p(x)\|^2 + (f'_0(x), p(x)) \\ & \leq \frac{1}{2} \|\bar{p}\|^2 + (f'_0(x), \bar{p}) + \sum_{i \in \mathcal{J}_\delta(x)} u^i(x) ((f'_i(x), \bar{p}) + f_i(x)) \\ & \leq \frac{1}{2} \|\bar{p}\|^2 + (f'_0(x), \bar{p}) + \sum_{i \in \mathcal{J}_\delta(x)} u^i(x) f_i(\bar{x}) \\ & \leq \frac{1}{2} \|\bar{p}\|^2 + (f'_0(x), \bar{p}) + u^i(x) f_i(\bar{x}). \end{aligned}$$

Hence $(f_i(\bar{x}) < 0!)$

$$u^i(x) \leq \frac{\left[\frac{1}{2} \|p(x)\|^2 + (f'_0(x), p(x)) \right] - \left[\frac{1}{2} \|\bar{p}\|^2 + (f'_0(x), \bar{p}) \right]}{f_i(\bar{x})} \quad (5.24)$$

where the numerator on the right-hand side of (5.24) contains a non-positive quantity since $p(x)$ is the solution of problem (5.4) and $\bar{p}$ satisfies the constraints of (5.4).

We shall show now that the right-hand side of (5.24) is bounded in Ω_0 . Indeed, since functions $f_i(x)$ are continuously differentiable, the quantity

$$\frac{1}{2} \|\bar{p}\|^2 + (f'_0(x), \bar{p}) = \frac{1}{2} \|\bar{x} - x\|^2 + (f'_0(x), \bar{x} - x)$$

is bounded in compact region Ω_0 . Therefore, the smaller quantity

$$\frac{1}{2} \|p(x)\|^2 + (f'_0(x), p(x))$$

has an upper bound. As to its lower bound we have

$$\begin{aligned} \frac{1}{2} \|p(x)\|^2 + (f'_0(x), p(x)) & \geq \frac{1}{2} \|p(x)\|^2 - \|f'_0(x)\| \|p(x)\| \\ & \geq -\frac{1}{2} \|f'_0(x)\|^2, \end{aligned}$$

i.e. with $x \in \Omega_0$ the quantity under consideration also has a lower bound.

Thus we have shown that in Ω_0 the right-hand sides of (5.24) have upper bounds, i.e. $u^i(x) \leq M$, $x \in \Omega_0$. The statement of the theorem directly follows from this fact.

Thus if the primal problem was a problem of convex programming, then any $\delta > 0$ suits the algorithm, provided the admissible region contains an interior point.

Some Generalizations

At the beginning of this section it was stated that if there are equality constraints, i.e. if the constraints are of the form (5.1), then the problem is reduced to the form (5.2) by substituting two inequalities for each equality.

Thus, the algorithm can be applied to the general problem (5.1) too. It should only be taken into account that if with a certain x we have

$$f_i(x) \geq F(x) - \delta \text{ and } -f_i(x) \geq F(x) - \delta$$

where $i \in \mathcal{J}_0$, then the system (5.4) comprises two inequalities

$$(f'_i(x), p) + f_i(x) \leq 0, \quad -(f'_i(x), p) - f_i(x) \leq 0 \quad (5.25)$$

which are equivalent to one equality

$$(f'_i(x), p) + f_i(x) = 0. \quad (5.26)$$

Therefore it is expedient to substitute in (5.4) one equality (5.26) for each pair of inequalities of the type (5.25). In passing to the dual problem this will lead to the corresponding multiplier u^i having an arbitrary sign which however does not impede the possibility of applying the algorithm of conjugate gradients (the subsection on p. 160).

Suppose now that in the primal problem in addition to constraints (5.2), there is a constraint imposed by the condition that point x belongs to a set X of simple structure. In this case it is expedient that the approximations obtained should lie in set X . We shall describe now how the algorithm is to be modified in this case. As we did previously we shall consider, without loss of generality, only the case with inequality constraints.

Thus let it be required to minimize $f_0(x)$, $x \in E^n$ with constraints

$$f_i(x) \leq 0, \quad i \in \mathcal{J}, \quad x \in X \quad (5.27)$$

where $\mathcal{J}$ is a finite set of indices and X is a convex closed set. It is assumed that there is an index i such that $f_i(x) \equiv 0$.

Suppose that there are constants $N > 0$ and $\delta > 0$ such that the following conditions are fulfilled:

(a) set

$$\Omega_N = \{x: f_0(x) + NF(x) \leq C_0, x \in X\},$$

$$C_0 = f_0(x_0) + NF(x_0),$$

is bounded and the initial approximation x_0 belongs to X ;

(b) the gradients of functions $f_i(x)$, $i \in \{0\} \cup \mathcal{J}$ in Ω_N satisfy Lipschitz' condition, i.e.

$$\|f'_i(x_1) - f'_i(x_2)\| \leq L \|x_1 - x_2\|;$$

(c) the problem

$$\min (f'_0(x), p) + \frac{1}{2} \|p\|^2, \\ (f'_i(x), p) + f_i(x) \leq 0, \quad i \in \mathcal{J}_\delta(x), \quad x + p \in X, \quad (5.28)$$

is solvable for p with any $x \in \Omega_N$ and there are Lagrange multipliers $u^i(x)$, $i \in \mathcal{J}_\delta(x)$ such that

$$\sum_{i \in \mathcal{J}_\delta(x)} u^i(x) \leq N.$$

Remark. Recall that Lagrange multipliers for problem (5.28) are all nonnegative numbers such that the following conditions are satisfied:

$$(f'_0(x), p(x)) + (p(x), p(x)) + \sum_{i \in \mathcal{J}_\delta(x)} u^i(x) [(f'_i(x), p(x)) + f_i(x)] \\ \leq (f'_0(x), p) + (p(x), p) + \sum_{i \in \mathcal{J}_\delta(x)} u^i(x) [(f'_i(x), p) + f_i(x)] \quad (5.29)$$

for all p such that

$$x + p \in X. \quad (5.30)$$

Besides

$$u^i(x) [(f'_i(x), p(x)) + f_i(x)] = 0, \quad i \in \mathcal{J}_\delta(x). \quad (5.31)$$

Thus condition (c) implies that not only the subsidiary problem (5.28) is solvable, but also that the minimum point $p = p(x)$ satisfies the necessary and sufficient conditions required by Kuhn-Tuckers' theorem.

The algorithm for solving problem (5.27) is constructed now as it was expounded in the subsection on p. 190. Only we take now as p_k vector $p(x_k)$ which is the solution of the new subsidiary problem (5.28).

We shall show that the algorithm is convergent, i.e. that the conclusions of theorem 5.1 hold and also that $x_k \in X$ with all k . It follows from the last assertion that any limit point of sequence $\{x_k\}$ lies in X . Since the proof of convergence differs from the proof of theorem 5.1 only in some details, there is no need in giving this proof completely. We shall point out only the main specific details.

First, since $x_k + p_k \in X$ and X is convex, we have $x_k + \alpha p_k \in X$ with all α lying between 0 and 1. Therefore if $x_k \in X$, then $x_{k+1} \in X$ too. And since $x_0 \in X$, by assumption, the whole sequence $\{x_k\}_{k=0}^\infty$ lies in X . Secondly from (5.29)-(5.31) with $p = 0$ we obtain that

$$(f'_0(x), p(x)) + \|p(x)\|^2 \leq \sum_{i \in \mathcal{J}_\delta(x)} u^i(x) f_i(x),$$

i.e.

$$(f'_0(x), p(x)) \leq \sum_{i \in \mathcal{J}_\delta(x)} u^i(x) f_i(x) - \|p(x)\|^2. \quad (5.32)$$

This inequality should be substituted for expression (5.7) which was used in obtaining estimate (5.13). All other calculations made in obtaining the estimates remain unchanged.

Finally, if at point x_* we have $p(x_*) = 0$, then it follows from (5.29)-(5.31) that conditions

$$\begin{aligned} (f'_0(x_*), p) + \sum_{i \in \mathcal{J}_\delta(x_*)} u^i(x_*) (f'_i(x_*), p) &\geq 0, \\ x_* + p &\in X, \quad u^i(x_*) f_i(x_*) = 0, \quad i \in \mathcal{J}_\delta(x_*) \end{aligned} \quad (5.33)$$

are fulfilled.

Besides, in this case it follows from (5.28) that

$$f_i(x_*) \leq 0, \quad i \in \mathcal{J}_\delta(x_*), \quad x_* \in X,$$

and it is also obvious that

$$f_i(x_*) < 0, \quad i \notin \mathcal{J}_\delta(x_*).$$

Thus point x_* satisfies all constraints (5.27) and (5.33) show that at this point the necessary conditions for an extremum are fulfilled.

So we have shown, as we did above, that if $p(x_*) = 0$, then at point x_* the necessary conditions for an extremum are satisfied. It is easy to show that the converse also holds, i.e. that the condition $p(x) = 0$ is necessary and sufficient for expecting point x to be an extremum point.

The proof of every limit point x_* of sequence $\{x_k\}$, $k = 0, 1, \dots$ satisfying the necessary conditions for an extremum is performed just in the same way as was used in proving theorem 5.1, i.e. by taking the limit in passing from relations (5.29)-(5.31) satisfied at points x_k to relations (5.33) satisfied at the limit point.

Problem of Linear Programming

Let now all functions $f_0(x)$, $f_i(x)$, $i \in \mathcal{J}$ in problem (5.2) be linear. We have then the problem of linear programming. Though the algorithm described is mostly important for the nonlinear case, its application to the problem of linear programming is also of avail. In particular, if set $\mathcal{J}$ comprises a great number of indices, then the problem of linear programming is one with many constraints. At the same time, with a small δ the subsidiary problem (5.4) has but a small number of constraints so that the general problem is reduced to the solving of a series of simpler problems. Besides as distinct from the simplex method, the method proposed does not accumulate

computation errors as it does not transform the original matrix of constraints from step to step.

For the problem of linear programming the conditions (a) and (c) (condition (b) is satisfied automatically) of the basic assumption are too strict for the convergency of the algorithm. We shall not dwell on the conditions of convergency for the problem of linear programming since our main purpose is to obtain an algorithm for the nonlinear case. It will be shown below that if the assumptions (a) and (c) for the problem of linear programming hold, then the algorithm converges after a finite number of steps. This fact characterizes to some extent the rate of convergence of the algorithm.

Theorem 5.3. *Let assumptions (a), (c) of the subsection (p. 189) hold and all functions $f_0(x)$, $f_i(x)$ which define problem (5.2) have the form*

$$f_i(x) = (a_i, x) - b_i.$$

Then the algorithm of the subsection on p. 190 converges after a finite number of steps.

Proof. Note from the beginning that in the case under consideration step α_k is equal to unity for sufficiently great k . Indeed, since all $f_i(x)$ are linear, Lipschitz' constant L is zero. It follows therefore from the formula for $\bar{\alpha}_k$ on p. 190 that

$$\begin{aligned} \bar{\alpha}_k &= \min \left(1, \frac{\delta}{F(x_k) + K \|p_k\|}, \frac{1-\varepsilon}{(N+1)L} \right) \\ &= \min \left(1, \frac{\delta}{F(x_k) + K \|p_k\|} \right). \end{aligned} \quad (5.34)$$

But it was proved above that $F(x) \rightarrow 0$, $\|p_k\| \rightarrow 0$. Therefore for sufficiently great k we have $\frac{\delta}{F(x_k) + K \|p_k\|} \geq 1$ and $\bar{\alpha}_k = 1$.

But the construction of $\bar{\alpha}_k$ is such that inequality (5.15) is fulfilled with $\alpha = \bar{\alpha}_k$. Since at each iteration the choice of α_k begins with halving $\alpha = 1$, it follows that inequality (5.5) which determines the choice of α_k will be satisfied immediately without any additional halvings, and step α_k will be just equal to 1.

Let point x_* be now a limit point of sequence $\{x_k\}$ generated by the algorithm. As we already know, this point is the solution of problem (5.2) for it satisfies all the constraints of the problem and, by theorem 5.1, the necessary conditions for a minimum as well; these conditions in our problem of linear programming are also sufficient.

We set

$$\mathcal{J}_0(x_*) = \{i \in \mathcal{J} : f_i(x_*) = 0\}. \quad (5.35)$$

Then $f_i(x_*) < 0$ for $i \in \mathcal{J}_0(x_*)$ so that

$$\varepsilon_0 = \max_{i \in \mathcal{J}_0(x_*)} f_i(x_*) < 0. \quad (5.36)$$

To simplify the notations in what follows we shall assume, without loss of generality, that all the sequence $\{x_k\}$ converges to x_* .

We consider now the subsidiary problem (5.4) at points of $\{x_k\}$:

$$\begin{aligned} \min (f'_0(x_k), p) + \frac{1}{2} \|p\|^2, \\ (f'_i(x_k), p) + f_i(x_k) \leq 0, \quad i \in \mathcal{J}_\delta(x_k). \end{aligned} \quad (5.37)$$

Its solution is $p_k = p(x_k)$. We denote the corresponding Lagrange's multipliers by u_k^i , $k \in \mathcal{J}_\delta(x_k)$ so that

$$u_k^i [(f'_i(x_k), p_k) + f_i(x_k)] = 0. \quad (5.38)$$

Let us show now that $\mathcal{J}_0(x_*) \subset \mathcal{J}_\delta(x_k)$ for all sufficiently great k . Indeed, if $i \in \mathcal{J}_\delta(x_k)$, then

$$f_i(x_k) < F(x_k) - \delta$$

and taking the limit with respect to k and since $F(x_k) \rightarrow 0$, we obtain that $f_i(x_*) \leq -\delta$ which would contradict the fact that $i \in \mathcal{J}_0(x_*)$.

Further, we introduce the notation

$$\tilde{\mathcal{J}}(x_k) = \{i \in \mathcal{J}_\delta(x_k) : u_k^i > 0\}.$$

We assert next that for great indices k ,

$$\tilde{\mathcal{J}}(x_k) \subset \mathcal{J}_0(x_*). \quad (5.39)$$

Indeed, if $i \in \mathcal{J}_0(x_*)$, then $f_i(x_*) \leq \varepsilon_0$. Since $p_k \rightarrow 0$ and $f'_i(x_k)$ are bounded and $x_k \rightarrow x_*$, we have with great k

$$|(f'_i(x_k), p_k)| \leq -\frac{\varepsilon_0}{4}, \quad f_i(x_k) \leq \frac{\varepsilon_0}{2},$$

and therefore

$$(f'_i(x_k), p_k) + f_i(x_k) \leq \frac{\varepsilon_0}{4} < 0.$$

Therefore if $u_k^i > 0$, then

$$u_k^i [(f'_i(x_k), p_k) + f_i(x_k)] < 0$$

but this contradicts (5.38).

Remark. In the argument no use was made of the linearity of $f_i(x)$ and therefore the statements that $\mathcal{J}_0(x_*) \subset \mathcal{J}_\delta(x_k)$ and $\tilde{\mathcal{J}}(x_k) \subset \mathcal{J}_0(x_*)$ hold in the general case of the nonlinear problem. These statements will be used in what follows.

As was shown in the subsection on p. 194 the dual of the subsidiary problem (5.37) is the maximization of function (5.21) with the constraints $u^i \geq 0$, $i \in \mathcal{J}_\delta(x_k)$. The Lagrange multipliers u_k^i are the solution of the dual problem and the equality of the optimum values in the original and dual problems holds, i.e. we have

$$\begin{aligned} (f'_0(x_k), p_k) + \frac{1}{2} \|p_k\|^2 \\ = -\frac{1}{2} \left\| f'_0(x_k) + \sum_{i \in \mathcal{J}_\delta(x_k)} u_k^i f'_i(x_k) \right\|^2 + \sum_{i \in \mathcal{J}_\delta(x_k)} u_k^i f_i(x_k). \end{aligned}$$

Since $p_k \rightarrow 0$, the left-hand side of the last expression tends to zero and consequently

$$-\frac{1}{2} \left\| f'_0(x_k) + \sum_{i \in \mathcal{J}_\delta(x_k)} u_k^i f'_i(x_k) \right\|^2 + \sum_{i \in \mathcal{J}_\delta(x_k)} u_k^i f_i(x_k) \rightarrow 0. \quad (5.40)$$

Note now that $u_k^i > 0$ only if $i \in \tilde{\mathcal{J}}(x_k)$. Besides,

$$f_i(x) = (a_i, x) - b_i, \quad i \in \{0\} \cup \mathcal{J},$$

so that $f'_i(x) = a_i$ and is independent of x . Therefore, (5.40) can be rewritten in the following form:

$$-\frac{1}{2} \left\| a_0 + \sum_{i \in \tilde{\mathcal{J}}(x_k)} u_k^i a_i \right\|^2 + \sum_{i \in \tilde{\mathcal{J}}(x_k)} u_k^i f_i(x_k) \rightarrow 0.$$

But $\tilde{\mathcal{J}}(x_k) \subset \mathcal{J}_0(x_*)$ as shown above and therefore $f_i(x_k) \rightarrow f_i(x_*) = 0$, since $f_i(x_*) = 0$ with $i \in \mathcal{J}_0(x_*)$, by definition. Therefore

$$-\frac{1}{2} \left\| a_0 + \sum_{i \in \tilde{\mathcal{J}}(x_k)} u_k^i a_i \right\|^2 \rightarrow 0.$$

But

$$\begin{aligned} -\frac{1}{2} \left\| a_0 + \sum_{i \in \tilde{\mathcal{J}}(x_k)} u_k^i a_i \right\|^2 \\ \leq \max_{u^i \geq 0, i \in \tilde{\mathcal{J}}(x_k)} -\frac{1}{2} \left\| a_0 + \sum_{i \in \tilde{\mathcal{J}}(x_k)} u^i a_i \right\|^2 \leq 0. \end{aligned} \quad (5.41)$$

We introduce the notation

$$\omega(\mathcal{Y}) = \max_{u^i \geq 0, i \in \mathcal{Y}} -\left\| a_0 + \sum_{i \in \mathcal{Y}} u^i a_i \right\|^2;$$

$\omega(\mathcal{Y})$ is a function defined in the set of indices $\mathcal{Y}$, $\mathcal{Y} \subset \mathcal{J}$. Since $\mathcal{Y} \subset \mathcal{J}$, this function can take only a finite number of values. It

follows from (5.41) that

$$\omega(\tilde{\mathcal{J}}(x_k)) \rightarrow 0.$$

But this means that $\omega(\tilde{\mathcal{J}}(x_k)) = 0$ for all sufficiently great k since as was just mentioned, $\omega(\mathcal{J})$ can take only a finite number of values.

Thus, for great k

$$\omega(\tilde{\mathcal{J}}(x_k)) = 0. \quad (5.42)$$

We now choose k so great that $\alpha_k = 1$; condition (5.42) is fulfilled and $\tilde{\mathcal{J}}(x_k) \subset \mathcal{J}_0(x_*)$. As $\alpha_k = 1$, we have $x_{k+1} = x_k + p_k$. Since $x_k \rightarrow x_*$, $p_k \rightarrow 0$, we can take that

$$f_i(x_{k+1}) \leq \frac{\varepsilon_0}{2} < 0, \quad i \in \mathcal{J}_0(x_*). \quad (5.43)$$

Let us consider again subsidiary problem (5.37). As p_k satisfies the constraints of (5.37) and $f_i(x)$ are linear, we have

$$f_i(x_{k+1}) = (f'_i(x_k), p_k) + f_i(x_k) \leq 0 \quad (5.44)$$

for $i \in \mathcal{J}_\delta(x_k)$ and consequently for $i \in \mathcal{J}_0(x_*)$ too as $\mathcal{J}_0(x_*) \subset \mathcal{J}_\delta(x_k)$. We have thus shown that x_{k+1} satisfies all constraints of problem (5.2).

We demonstrate that x_{k+1} is really the solution of problem (5.2). Indeed, it follows from (5.38) and the definition of set $\tilde{\mathcal{J}}(x_k)$ that

$$f_i(x_{k+1}) = 0, \quad i \in \tilde{\mathcal{J}}(x_k). \quad (5.45)$$

But (5.42) means that there are numbers $u_0^i \geq 0$, $i \in \tilde{\mathcal{J}}(x_k)$ such that

$$a_0 + \sum_{i \in \tilde{\mathcal{J}}(x_k)} u_0^i a_i = 0. \quad (5.46)$$

Setting now $u_0^i = 0$, $i \notin \tilde{\mathcal{J}}(x_k)$ we obtain that there are numbers $u_0^i \geq 0$ such that conditions

$$a_0 + \sum_{i \in \mathcal{J}} u_0^i a_i = 0, \quad u_0^i f_i(x_{k+1}) = 0,$$

are fulfilled.

But the last relations (see Chap. I, Sec. 3) are the necessary and sufficient conditions for point x_{k+1} to be the solution of the problem of linear programming.

Thus the algorithm provides for the solution after a finite number of steps. Q.E.D.

Local Estimate of the Rate of Convergence

It was shown in the preceding subsection that the algorithm proposed converges after a finite number of steps in the linear case. Here we shall show that in the general nonlinear case the algorithm converges at a geometrical rate and with certain favourable circumstances even at a quadratic rate.

Theorem 5.4. *Let x_* be the solution of problem 5.2 and the following conditions hold:*

(a) *For any sufficiently small $\delta > 0$, the subsidiary problem (5.4) is solvable.*

(b) *Functions $f_i(x)$ are twice continuously differentiable and the gradients $f'_i(x_*)$, $i \in \mathcal{J}_0(x_*)$, where*

$$\mathcal{J}_0(x_*) = \{i: f_i(x_*) = 0, i \in \mathcal{J}\},$$

are linearly independent.

(c) *At point x_* the necessary condition for a minimum is satisfied in the form*

$$f'_0(x_*) + \sum_{i \in \mathcal{J}_0(x_*)} u_0^i f'_i(x_*) = 0$$

and $u_0^i > 0$, $i \in \mathcal{J}_0(x_*)$.

(d) *The sufficient condition for a local minimum, i.e.*

$$(p, L''(x_*, u_0)p) > 0,$$

holds for all $p \neq 0$ which also satisfy the condition

$$(p, f'_i(x_*)) = 0, i \in \mathcal{J}_0(x_*)$$

where

$$L(x, u) = f_0(x) + \sum_{i \in \mathcal{J}_0(x_*)} u^i f_i(x)$$

and L'' is a matrix of second derivatives of $L(x, u)$ with respect to x . Then there is a neighbourhood Ω of point x_* , $\delta_0 > 0$ and $\alpha > 0$ such that the process

$$x_{k+1} = x_k + \alpha p_k \tag{5.47}$$

converges to point x_* from any initial approximation $x_0 \in \Omega$ at a geometric rate, i.e. there is a number $0 \leq q < 1$ such that $\|x_* - x_k\| \leq Cq^k$ for all sufficiently great k .

Proof. The basic idea of the proof is as follows. As was shown above, at point x_* the equation

$$p(x_*) = 0$$

is satisfied.

Process (5.47) is a simple iterative process for solving the last equation. Therefore in order to estimate the rate of convergence we can use Ostrowski's theorem which is formulated below. This theorem requires an estimate of the eigenvalues of the matrix of first derivatives of $p(x)$ at point x_* . Therefore the main problem will be the calculation of this matrix and its eigenvalues.

We shall break the proof of the theorem into several parts.

We take

$$\mathcal{J}_0(x_*) = \{i \in \mathcal{J} : f_i(x_*) = 0\},$$

$$\varepsilon_0 = \max_{i \notin \mathcal{J}_0(x_*)} f_i(x_*) < 0.$$

Lemma 5.2. *Let the conditions of the theorem be fulfilled and let $\delta < -\frac{\varepsilon_0}{2}$. Then there is a neighbourhood of point x_* such that $\mathcal{J}_\delta(x) = \mathcal{J}_0(x_*)$ and $p(x)$ is continuously differentiable with respect to x in this neighbourhood. Moreover, the set*

$$\tilde{\mathcal{J}}(x) = \{i \in \mathcal{J}_\delta(x) : (f'_i(x), p(x)) + f_i(x) = 0\}$$

coincides with set $\mathcal{J}_0(x_*)$.

Proof. Since all functions $f_i(x)$ are continuous, there is a neighbourhood of point x_* such that

$$-\frac{\delta}{2} < f_i(x) < \frac{\delta}{2}, \quad i \in \mathcal{J}_0(x_*), \quad (5.48)$$

$$f_i(x) < \frac{\varepsilon_0}{2}, \quad i \notin \mathcal{J}_0(x_*). \quad (5.49)$$

Recall now that

$$F(x) = \max \{0, \max_{i \in \mathcal{J}} f_i(x)\}$$

and $i \in \mathcal{J}_\delta(x)$ if $f_i(x) \geq F(x) - \delta$. It follows from (5.48), (5.49) that

$$0 \leq F(x) < \frac{\delta}{2} \quad (5.50)$$

and if $i \notin \mathcal{J}_0(x_*)$, then

$$F(x) - \delta \geq -\delta > \frac{\varepsilon_0}{2} > f_i(x),$$

i.e. $i \notin \mathcal{J}_\delta(x)$. On the other hand, if $i \in \mathcal{J}_0(x_*)$ then it follows from (5.50) that $F(x) - \delta < -\frac{\delta}{2}$ and consequently

$$f_i(x) > -\frac{\delta}{2} > F(x) - \delta,$$

i.e. $i \in \mathcal{J}_\delta(x)$.

Thus we have shown that $\mathcal{J}_\delta(x) = \mathcal{J}_0(x_*)$ in a certain neighbourhood of x_* .

Recall now that if $p(x)$ is the solution of problem (5.4), then conditions (5.6) hold; these conditions can be rewritten in the equivalent form:

$$p(x) + f'_0(x) + \sum_{i \in \tilde{\mathcal{J}}(x)} u^i(x) f'_i(x) = 0, \quad (5.51.1)$$

$$(f'_i(x), p(x)) + f_i(x) = 0, \quad i \in \tilde{\mathcal{J}}(x), \quad (5.51.2)$$

$$(f'_i(x), p(x)) + f_i(x) < 0, \quad i \notin \tilde{\mathcal{J}}(x), \quad i \in \mathcal{J}_\delta(x) \quad (5.51.3)$$

where $u^i(x) \geq 0$.

We introduce the following notations: $\mathcal{Y} \subset \mathcal{J}_\delta(x) (= \mathcal{J}_0(x_*))$; $f'_{\mathcal{Y}}(x)$ for a matrix whose rows are $f'_i(x)$, $i \in \mathcal{Y}$; $f_{\mathcal{Y}}(x)$ for a column-vector with components $f_i(x)$, $i \in \mathcal{Y}$ and $u_{\mathcal{Y}}$ for a column-vector with components u^i , $i \in \mathcal{Y}$. Then equations (5.51.1), (5.51.2) can be rewritten as follows:

$$\begin{aligned} p(x) + f'_0(x) + f'^*_{\mathcal{Y}}(x) u_{\mathcal{Y}}(x) &= 0, \\ f'_{\mathcal{Y}}(x) p(x) + f_{\mathcal{Y}}(x) &= 0, \quad \mathcal{Y} = \tilde{\mathcal{J}}(x). \end{aligned} \quad (5.52)$$

The last expressions can be considered to be a linear system of equations to be solved for $p(x)$ and $u_{\mathcal{Y}}(x)$. It is easy to see that in a certain neighbourhood of x_* , system (5.52) has only one solution expressed by the formulas

$$\begin{aligned} u_{\mathcal{Y}}(x) &= (f'_{\mathcal{Y}}(x) f'^*_{\mathcal{Y}}(x))^{-1} [f_{\mathcal{Y}}(x) - f'_{\mathcal{Y}}(x) f'_0(x)], \\ p(x) &= -f'_0(x) - f'^*_{\mathcal{Y}}(x) u_{\mathcal{Y}}(x). \end{aligned} \quad (5.53)$$

It follows from these formulas that if set $\mathcal{Y}$ is fixed, then $u_{\mathcal{Y}}(x)$ and $p(x)$ are continuously dependent on x .

Let now $x_k \rightarrow x_*$. We shall show that for all great k

$$\tilde{\mathcal{J}}(x_k) = \mathcal{J}_0(x_*).$$

Suppose that our statement is not fulfilled and there are great numbers k such that $\tilde{\mathcal{J}}(x_k)$ is a subset of $\mathcal{J}_0(x_*)$. Since there can be but a finite number of different sets $\tilde{\mathcal{J}}(x)$, we can take, without loss of generality, that a sequence $x_k \rightarrow x_*$ is chosen such that $\tilde{\mathcal{J}}(x_k) = \mathcal{Y}$, $\mathcal{Y} \subset \mathcal{J}_0(x_*)$.

Substituting now x_k for x in (5.51) and taking the limit ($p(x_k) \rightarrow \tilde{p}$, $u^i(x_k) \rightarrow \tilde{u}^i$, $i \in \mathcal{Y}$) we obtain that

$$\tilde{p} + f'_0(x_*) + \sum_{i \in \mathcal{Y}} \tilde{u}^i f'_i(x_*) = 0,$$

$$(f'_i(x_*), \tilde{p}) + f_i(x_*) = 0, \quad i \in \mathcal{Y},$$

$$(f'_i(x_*), \tilde{p}) + f_i(x_*) \leq 0, \quad i \notin \mathcal{Y}, \quad i \in \mathcal{J}_0(x_*) = \mathcal{J}_\delta(x_*)$$

where $\tilde{u}^i \geq 0$, for $u^i(x_k) \geq 0$. But these last relations show that $\tilde{p}$ is the solution of the subsidiary problem (5.4) for point x_* , i.e. $\tilde{p} = p(x_*)$. But point x_* is the solution of problem (5.2) and therefore $p(x_*) = 0$. Consequently,

$$f'_0(x_*) + \sum_{i \in \mathcal{Y}} \tilde{u}^i f'_i(x_*) = 0.$$

Using now condition (c) of the theorem we obtain from the last expression that

$$\sum_{i \in \mathcal{Y}} (u'_0 - \tilde{u}^i) f'_i(x_*) + \sum_{i \in \mathcal{J}_0(x_0) \setminus \mathcal{Y}} u'_0 f'_i(x_*) = 0$$

and this contradicts condition (b) of the theorem. Thus, in a certain neighbourhood of point x_* , set $\tilde{\mathcal{J}}(x)$ coincides with $\mathcal{J}_0(x_*)$. From $\tilde{\mathcal{J}}(x)$ being constant and formulas (5.53) it follows directly that $u_{\mathcal{Y}}(x)$ and $p(x)$, $\mathcal{Y} = \mathcal{Y}_0(x_*)$ are continuously differentiable with respect to x since, by condition (b), $f_i(x)$ are twice continuously differentiable.

Remark. Thus, in a small neighbourhood of x_* , $p(x)$ and $u_{\mathcal{Y}}(x)$ is the solution of the system of equations (5.52) with a constant set $\mathcal{Y} = \mathcal{J}_0(x_*)$. Therefore we shall omit index $\mathcal{Y}$ in $u_{\mathcal{Y}}(x)$.

Lemma 5.3. *The matrix $p'(x)$ of derivatives of vector $p(x)$, i.e. the matrix with elements $\partial p^i(x)/\partial x^j$, $i, j = 1, \dots, n$, where $p^i(x)$ is the i -th component of vector $p(x)$, at point x_* has the following form:*

$$p'(x_*) = -[P + (I - P)L''(x_*, u_0)]$$

where

$$P = f'^*_{\mathcal{J}_0(x_*)}(x_*) (f'_{\mathcal{J}_0(x_*)}(x_*) f'^*_{\mathcal{J}_0(x_*)}(x_*))^{-1} f'_{\mathcal{J}_0(x_*)}(x_*)$$

and $u_0 = u(x_*)$.

Proof. By differentiating the first of formulas (5.52) we obtain

$$p'(x_*) = -L''(x_*, u_0) - \sum_{j \in \mathcal{J}_0(x_*)} f'_j(x_*) (u^{j'}(x_*))^* \quad (5.54)$$

where

$$u^{j'}(x) = \begin{pmatrix} \frac{\partial u^j(x)}{\partial x^1} \\ \vdots \\ \frac{\partial u^j(x)}{\partial x^n} \end{pmatrix}.$$

From formula (5.51.2) by differentiating $(p(x_*) = 0)$, we obtain

$$f_i^{*'}(x_*) p'(x_*) + f_i^{*'}(x_*) = 0, \quad i \in \mathcal{J}_0(x_*) \quad (5.55)$$

Note now that operator P defined in formulating the lemma is the operator of projecting on the subspace spanned by the vectors $f_i'(x_*)$, $i \in \mathcal{J}_0(x_*)$. Indeed, this can be seen (see also the subsection on p. 147) from easily verified relations:

- (1) $P f_{\mathcal{J}_0(x_*)}^{*'}(x_*) = f_{\mathcal{J}_0(x_*)}^{*'}(x_*)$ or $P f_i'(x_*) = f_i'(x_*)$, $i \in \mathcal{J}_0(x)$;
- (2) $P^* = P$, $P^2 = P$;
- (3) $(I - P)P = 0$.

If we now rewrite (5.55) in the form

$$f_{\mathcal{J}_0(x_*)}^{*'}(x_*) p'(x_*) + f_{\mathcal{J}_0(x_*)}^{*'}(x_*) = 0,$$

then taking into account the expression for P , we obtain

$$P p'(x_*) = -P. \quad (5.56)$$

Further, it follows from relation (1) for P that $(I - P) f_i'(x_*) = 0$. Therefore applying $(I - P)$ to both sides of (5.54), we obtain

$$(I - P) p'(x_*) = -(I - P) L''(x_*, u_0). \quad (5.57)$$

Adding (5.56) and (5.57) we obtain the required formula for $p'(x_*)$.

Lemma 5.4. *Eigenvalues γ_j of matrix $p'(x_*)$ can be characterized as follows: $\gamma_j = -1$ for $j = 1, 2, \dots, m$, where $m \leq n$ is the number of indices in set $\mathcal{J}_0(x_*)$. $\gamma_j = -\lambda_{j-m}$, $j = m + 1, \dots, n$, where λ_j , $j = 1, \dots, n - m$, are the eigenvalues of the matrix $(I - P) L''(x_*, u_0) (I - P)$ and $\lambda_j > 0$, $j = 1, \dots, n - m$.*

Proof. Let σ be the eigenvalue and y the eigenvector of matrix $p'(x_*)$. Then, according to lemma 5.3, we have

$$-Py - (I - P) L''(x_*, u_0) y = \sigma y = \sigma Py + \sigma (I - P) y.$$

We use now the relation $P(I - P) = 0$ and by multiplying the last equality in turn by P and $I - P$ we obtain:

$$-Py = \sigma Py, \quad (5.58)$$

$$-(I - P) L''(x_*, u_0) y = \sigma (I - P) y. \quad (5.59)$$

There are two possible cases.

(1) $Py \neq 0$. Then it follows from (5.58) that $\sigma = -1$.

(2) $Py = 0$. In this case, $(I - P)y = y$ and (5.59) can be rewritten as follows:

$$(I - P) L''(x_*, u_0) (I - P) y = -\sigma y, \quad (5.60)$$

i.e. σ is the eigenvalue of matrix $(I - P) L''(I - P)$. This matrix is symmetric since $P = P^*$ and $L'' = (L'')^*$ being the matrix of second derivatives of function L . Moreover, the matrix under consideration is nonnegative definite. Indeed, for any w we have

$$(w, (I - P) L''(I - P) w) = (z, L'' z)$$

where

$$z = (I - P) w.$$

But $f'_{\mathcal{J}_0(x_*)}(x_*) z = f'_{\mathcal{J}_0(x_*)}(x_*) (I - P) w = 0$ and therefore $(z, L'' z) \geq 0$ by the condition (d) of theorem 5.4, the equality sign being possible only if $z = (I - P) w = 0$. It follows from the symmetry of matrix $(I - P) L''(I - P)$ that its eigenvalues and eigenvectors are real. As $y \neq 0$ and $y = Py + (I - P)y$, it follows from $Py = 0$ that $(I - P)y \neq 0$, and therefore from (5.60) we obtain

$$-\sigma (y, y) = (y, (I - P) L''(I - P) y) = (y, L'' y) > 0.$$

Thus, $-\sigma \neq 0$ and consequently $\sigma = -\lambda_j$, where $\lambda_j > 0$ is the eigenvalue of matrix $(I - P) L''(I - P)$.

Thus we have proved that the eigenvalues of matrix $p'(x_*)$ are real and equal either to -1 or to $-\lambda_j$, $\lambda_j > 0$. It remains to determine only the number of eigenvalues which are equal to -1 .

Due to the fact that

$$Pf'_i(x_*) = f'_i(x_*), \quad i \in \mathcal{J}_0(x_*),$$

operator $(I - P)$ has m eigenvectors $f'_i(x_*)$ which correspond to the zero eigenvalue. Therefore, matrix $(I - P) L''(I - P)$ also has m zero eigenvalues. On the other hand, as we have seen, matrix $p'(x)$ has all the n nonzero eigenvalues, each being either equal to -1 or the one of matrix $(I - P) L''(I - P)$. Clearly, this is possible only if the statement of lemma 5.4 holds true.

We now complete the proof of theorem 5.4. It follows from Ostrowski's theorem that if x_* is the solution of the equation $p(x) = 0$ and the eigenvalues of matrix $I + \alpha p'(x_*)$ have a modulus less than unity, then the method of simple iteration $x_{k+1} = x_k + \alpha p(x_k)$ converges from all points of a certain neighbourhood of point x_* and at the same time the following rate of convergence holds: for each $\varepsilon > 0$ there is a number $C(\varepsilon)$ such that $\|x_* - x_k\| \leq C(\varepsilon)(q_0 + \varepsilon)^k$, where q_0 is the greatest of the moduli of the eigenvalues of matrix $I + \alpha p'(x_*)$.

Consider now the eigenvalues of matrix $I + \alpha p'(x_*)$. They are equal either to $1 - \alpha$ or to $1 - \alpha\lambda_j$. We choose now α so that all of the following inequalities be satisfied:

$$1 - \alpha > -1, \quad 1 - \alpha\lambda_j > -1, \quad j = 1, \dots, n - m,$$

i.e. that $0 < \alpha < \min \{2, 2/\lambda_0\}$, where $\lambda_0 = \max \lambda_j$, $j = 1, \dots, n - m$. Then all the eigenvalues of matrix $I + \alpha p'(x_*)$ will have moduli less than unity; hence, referring also to Ostrowski's results, we have that theorem 5.4 holds.

Theorem 5.5. *Let the conditions of the preceding theorem be satisfied and, besides, m (the number of indices in set $\mathcal{J}_0(x_*)$) be equal to n (the dimension of the space). In this case, process (5.47) converges from a certain neighbourhood of point x_* with $\alpha = 1$ at a quadratic rate.*

Proof. It follows from lemma 5.4 for the case under consideration that all the eigenvalues of matrix $p'(x_*)$ are equal to -1 , and therefore the eigenvalues of matrix $I + \alpha p'(x_*)$ are equal to $1 - \alpha$. If $\alpha = 1$, then all the eigenvalues are equal to zero and $q_0 = 0$. Therefore according to Ostrowski's theorem, we obtain $\|x_* - x_h\| \leq C(\varepsilon) \varepsilon^h$ and this means that the process converges at a higher rate than that of any geometric progression. In fact in this case, process (5.47) passes into Newton's method for solving systems of equations $f_i(x) = 0$, $i \in \mathcal{J}_0(x_*)$ which as well known and as shown below in Sec. 6 converges quadratically.

Remark. All the arguments in this subsection were conducted for the case of a problem with only inequality constraints. It is obvious, however, that all the results obtained can be applied to the case with equality constraints.

6. LINEARIZATION METHOD: SOLVING SYSTEMS OF EQUALITIES AND INEQUALITIES AND FINDING THE MINIMAX

In this section, the linearization method is applied to two problems closely connected to the usual problem of mathematical programming. It proves that in this case one can succeed in constructing effective algorithms which have a fast rate of convergence.

Systems of Equalities and Inequalities

Given two finite sets of indices $\mathcal{J}^-$ and $\mathcal{J}^0$ and functions $f_i(x)$, $x \in E^n$. To find the solution of the following system:

$$f_i(x) \leq 0, \quad i \in \mathcal{J}^-, \quad f_i(x) = 0, \quad i \in \mathcal{J}^0. \quad (6.1)$$

Suppose that functions $f_i(x)$ have continuous gradients $f'_i(x)$ and

also that the gradients satisfy Lipschitz' condition with constant L :

$$\| f'_i(x_1) - f'_i(x_2) \| \leq L \| x_1 - x_2 \|.$$

The norm of vectors is everywhere Euclidean.

We use the notation:

$$\begin{aligned} F(x) &= \max \left(\max_{i \in \mathcal{J}^-} f_i(x), \max_{i \in \mathcal{J}^0} |f_i(x)| \right), \\ \mathcal{J}_\delta^-(x) &= \{i: i \in \mathcal{J}^-, f_i(x) \geq F(x) - \delta\}, \\ \mathcal{J}_\delta^0(x) &= \{i: i \in \mathcal{J}^0, |f_i(x)| \geq F(x) - \delta\}. \end{aligned}$$

We choose an initial point x_0 and assume that for all x that satisfy the inequality $F(x) \leq F(x_0)$, the gradients $f'_i(x)$ are limited in norm by constant K .

Basic assumption. There are numbers $\delta > 0$ and $C > 0$ such that for all x for which $F(x) > 0$, $F(x) \leq F(x_0)$ the following system is solvable for p :

$$\begin{aligned} (f'_i(x), p) + f_i(x) &\leq 0, \quad i \in \mathcal{J}_\delta^-(x), \\ (f'_i(x), p) + f_i(x) &= 0, \quad i \in \mathcal{J}_\delta^0(x). \end{aligned} \tag{6.2}$$

Let $p(x)$ be the solution of (6.2) that has the minimum norm. Then for x such that $F(x) > 0$,

$$\| p(x) \| \leq CF(x). \tag{6.3}$$

The inequality (6.3) characterizes to a certain extent the regular solvability of system (6.2). In particular, if system (6.2) is transformed into a system of n equations in n unknowns, condition (6.3) is equivalent to the assumption that the matrix of the corresponding system is nonsingular. As will be shown further on, (6.3) holds if the gradients $f'_i(x)$, $i \in \mathcal{J}_\delta^-(x) \cup \mathcal{J}_\delta^0(x)$ are linearly independent for all x , $F(x) > 0$.

We turn now to the construction of the algorithm. The successive approximations are constructed by the formula

$$x_{k+1} = x_k + \alpha_k p_k, \quad p_k \equiv p(x_k) \tag{6.4}$$

where parameter α_k is chosen by sequentially halving unity until the following inequality is satisfied:

$$F(x_k + \alpha_k p_k) \leq (1 - \varepsilon \alpha_k) F(x_k) \tag{6.5}$$

where ε is any number, chosen from the beginning, $0 < \varepsilon < 1$. Clearly, formula (6.4) is applicable if $F(x) > 0$. Otherwise, the process stops and x_k is the solution of (6.1).

Convergence of the Algorithm

The implementing of the algorithm proposed is characterized by the following theorem.

Theorem 6.1. *Let all the assumptions of the preceding subsection be fulfilled. Then sequence $\{x_k\}$, $k = 0, 1, \dots$, generated by the algorithm according to formula (6.4) converges to $\bar{x}$, the solution of system (6.1), and at the same time*

- (a) *for a sufficiently great k , $\alpha_k = 1$;*
- (b) *for a sufficiently great k ,*

$$F(x_{k+1}) \leq LC^2 F^2(x_k);$$

- (c) *for any q , $0 < q < 1$ there is a number $k(q)$ such that*

$$\|\bar{x} - x_k\| \leq \frac{q^{2^k - k(q)}}{LC(1-q)} \quad (6.6)$$

for all $k \geq k(q)$.

Proof. Obviously, if $F(x_k) \leq 0$ at a certain step, all the statements are proved. Therefore we suppose that $F(x_k) > 0$ for all k .

First of all we show that the choice of α_k under condition (6.5) is always feasible. For $i \in \mathcal{J}_\delta^-(x_k)$, using Taylor's formula, we have:

$$\begin{aligned} f_i(x_k + \alpha p_k) &= f_i(x_k) + \alpha(f'_i(x_k + \theta_i \alpha p_k), p_k) \\ &= f_i(x_k) + \alpha(f'_i(x_k), p_k) + \alpha(f'_i(x_k + \theta_i \alpha p_k) - f'_i(x_k), p_k) \end{aligned}$$

where $0 \leq \theta_i \leq 1$. But since p_k satisfies (6.2), we have

$$(f'_i(x_k), p_k) \leq -f_i(x_k).$$

Further

$$\begin{aligned} (f'_i(x_k + \theta_i \alpha p_k) - f'_i(x_k), p_k) &\leq \|p_k\| \|f'_i(x_k + \theta_i \alpha p_k) - f'_i(x_k)\| \\ &\leq \|p_k\| \|\theta_i \alpha p_k\| L \leq \alpha L \|p_k\|^2. \end{aligned}$$

Therefore using (6.3), we obtain

$$\begin{aligned} f_i(x_k + \alpha p_k) &\leq f_i(x_k) - \alpha f_i(x_k) + \alpha^2 L \|p_k\|^2 \\ &\leq (1 - \alpha) F(x_k) + \alpha^2 LC^2 F^2(x_k). \end{aligned} \quad (6.7)$$

For $i \in \mathcal{J}^-$, $i \notin \mathcal{J}_\delta^-(x_k)$, $f_i(x_k) < F(x_k) - \delta$ and therefore

$$\begin{aligned} f_i(x_k + \alpha p_k) &= f_i(x_k) + \alpha(f'_i(x_k + \theta_i \alpha p_k), p_k) \\ &\leq F(x_k) - \delta + \alpha K \|p_k\| \leq F(x_k) - \delta + \alpha KCF(x_k). \end{aligned} \quad (6.8)$$

Quite similarly, we have for $i \in \mathcal{J}_\delta^0(x_k)$

$$|f_i(x_k + \alpha p_k)| \leq (1 - \alpha) F(x_k) + \alpha^2 LC^2 F^2(x_k) \quad (6.9)$$

and for $i \notin \mathcal{J}_\delta^0(x_k)$

$$|f_i(x_k + \alpha p_k)| \leq F(x_k) - \delta + \alpha KCF(x_k). \quad (6.10)$$

Note now that

$$(1 - \alpha) F(x_k) \geq F(x_k) - \delta + \alpha CK F(x_k)$$

if $\alpha \leq \alpha_k^1$, where

$$\alpha_k^1 = \frac{\delta}{(1 + CK) F(x_k)}.$$

Therefore for $\alpha \leq \alpha_k^1$ it follows from (6.7)-(6.10) that

$$F(x_k + \alpha p_k) \leq (1 - \alpha) F(x_k) + \alpha^2 LC^2 F^2(x_k)$$

or

$$F(x_k + \alpha p_k) \leq F(x_k) - \alpha F(x_k) [1 - \alpha LC^2 F(x_k)]. \quad (6.11)$$

If $\alpha \leq \alpha_k^2$, where

$$\alpha_k^2 = \frac{1 - \varepsilon}{LC^2 F(x_k)},$$

then $1 - \alpha LC^2 F(x_k) \geq \varepsilon$ and therefore (6.11) can be rewritten as

$$\begin{aligned} F(x_k + \alpha p_k) &\leq F(x_k) - \alpha \varepsilon F(x_k), \\ \alpha &\leq \min \{\alpha_k^1, \alpha_k^2\}. \end{aligned} \quad (6.12)$$

It is now clear that if we proceed to reduce α beginning with $\alpha = 1$, then inequality (6.12) will hold after a finite number of trials and the α_k chosen will satisfy the inequality

$$\alpha_k \geq \min \left\{ 1, \frac{1}{2} \alpha_k^1, \frac{1}{2} \alpha_k^2 \right\}. \quad (6.13)$$

Thus, we have proved that the choice of α_k under condition (6.5) is feasible and that this choice can be realized after a finite number of operations.

We show that $F(x_k) \rightarrow 0$. Indeed, it follows from (6.5) that $F(x_k)$ decreases monotonically. Therefore, it can be concluded from the formulas for α_k^1 and α_k^2 that these quantities increase with increasing k . Consequently, formula (6.13) permits us to conclude that $\alpha_k \geq \bar{\alpha} > 0$ and so

$$F(x_{k+1}) \leq (1 - \varepsilon \alpha_k) F(x_k) \leq (1 - \varepsilon \bar{\alpha}) F(x_k).$$

Therefore $F(x_k) \leq (1 - \varepsilon \bar{\alpha})^k F(x_0)$, hence $F(x_k) \rightarrow 0$. But then $\alpha_k^1 \rightarrow +\infty$, $\alpha_k^2 \rightarrow +\infty$ as can be seen directly from the formulas for these quantities. Therefore, (6.13) permits us to conclude that $\alpha_k = 1$ for a sufficiently great k . But (6.11) shows for all such k , if $\alpha = 1$ is substituted into it, that

$$F(x_{k+1}) \leq LC^2 F^2(x_k). \quad (6.14)$$

Thus statements (a) and (b) of the theorem have been proved.

We can now assert that there is a k_0 such that $\alpha_k = 1$ for $k \geq k_0$ and (6.14) is satisfied. Therefore, by (6.3),

$$\|x_{k+1} - x_k\| = \|p_k\| \leq CF(x_k).$$

We set $v_k = LC^2F(x_k)$. Then $v_k \rightarrow 0$ and (by (6.14)) $v_{k+1} \leq v_k^2$. Let q be such that $0 < q < 1$. Then there is a $k(q)$ such that $v_k < q$ for $k \geq k(q)$. Therefore $v_{k+1} \leq qv_k$, $k \geq k(q)$. Hence

$$v_k \leq v_{k(q)}^{2^{k-k(q)}} \leq q^{2^{k-k(q)}}, \quad v_m \leq q^{m-k} v_k, \quad m \geq k \geq k(q).$$

This permits us to obtain the following estimate:

$$\begin{aligned} \|x_m - x_k\| &\leq \sum_{j=k}^{m-1} \|x_{j+1} - x_j\| \leq \frac{1}{LC} \sum_{j=k}^{m-1} v_j \\ &\leq \frac{v_k}{LC} \sum_{j=0}^{m-1-k} q^j \leq \frac{v_k}{LC(1-q)} \leq \frac{q^{2^{k-k(q)}}}{LC(1-q)}. \end{aligned} \quad (6.15)$$

It follows from this estimate (according to the well known Cauchy criterion) that sequence $\{x_k\}$ converges to a certain point $\bar{x}$. Since $F(x_k) \rightarrow 0$, we have $F(\bar{x}) = 0$, i.e. $\bar{x}$ is the solution of system (6.1). Moreover, taking the limit in (6.15) as $m \rightarrow \infty$ we obtain

$$\|\bar{x} - x_k\| \leq \frac{q^{2^{k-k(q)}}}{LC(1-q)}.$$

Q.E.D.

Remarks

Remark 1. Let us be solving a system of n equations $f_i(x) = 0$, $i = 1, \dots, n$, where $x \in E^n$. Then

$$f_i(x) \geq F(x) - \delta, \quad i = 1, \dots, n,$$

$$F(x) = \max_{1 \leq i \leq n} |f_i(x)|$$

for any δ , provided x is sufficiently close to the solution $\bar{x}$. Therefore, $\mathcal{J}_\delta^0(x) = \{1, 2, \dots, n\}$ and system (6.2) takes the form

$$(f'_i(x), p) + f_i(x) = 0, \quad i = 1, \dots, n. \quad (6.16)$$

Therefore the method proposed coincides with Newton's method in which iterations are performed by the formula $x_{k+1} = x_k + p(x_k)$, where $p(x)$ is the solution of system (6.16). The condition for the convergence of Newton's method is the nonsingularity at point $\bar{x}$ of matrix $f'(\bar{x})$, where $f'(x)$ is an $n \times n$ matrix whose rows are $f'_i(x)$. In this case, $p(x) = -(f'(x))^{-1} f(x)$, where $f(x)$ is a column-vector whose components are $f_i(x)$. But it follows from the

last formula that

$$\|p(x)\| \leq \| (f'(x))^{-1} \| \|f(x)\| \leq C_0 \| (f'(x))^{-1} \| F(x)$$

where C_0 is a constant. It can be seen from this inequality that (6.3) holds in a certain neighbourhood of point $\bar{x}$.

Thus it follows from the theorem proved that the usual Newton's method is locally convergent in solving a system of n equations with n unknowns.

Remark 2. If only one equation $f(x) = 0$ in n unknowns is to be solved, then system (6.2) takes the form

$$(f'(x), p) + f(x) = 0 \quad (6.17)$$

and it is required to find the solution of this equation with a minimum norm, i.e. to find the minimum of $\|p\|^2$ with constraints (6.17). Using the rule of Lagrange multipliers, we have in this case

$$p(x) = -\frac{f(x)}{\|f'(x)\|^2} f'(x),$$

hence

$$\|p(x)\| = \frac{1}{\|f'(x)\|} |f(x)|.$$

Clearly, formula (6.3) will be satisfied if $\|f'(x)\| \geq \gamma$ for all x .

Remark 3. The finding of vector $p(x)$ at each step involves the solving of the problem of minimization of $\|p\|^2$ with constraints (6.2). This is a problem of quadratic programming. Concerning the methods of solving it, we can use the same information given in Sec. 5 about the solving of the subsidiary problem of quadratic programming which arises in the linearization method.

Sufficient Conditions of Convergence

The main condition (6.3) which guarantees the convergence of the algorithm is not easy to check. This subsection describes conditions that can be checked more effectively. In particular, for the convex case if there is an interior point in the domain defined by expressions (6.1), the conditions guarantee the convergence of the algorithm.

Let the system contain only inequality constraints, i.e.

$$f_i(x) \leq 0, \quad i \in \mathcal{J}^-. \quad (6.18)$$

Then the subsidiary system (6.2) takes the form

$$(f'_i(x), p) + f_i(x) \leq 0, \quad i \in \mathcal{J}_\delta^-(x). \quad (6.19)$$

Clearly, this system can be solved with $F(x) > 0$ if the system

$$(f'_i(x), p) + F(x) \leq 0, \quad i \in \mathcal{J}_\delta^-(x) \quad (6.20)$$

is solvable.

Lemma 6.1. *If $F(x) > 0$, then system (6.20) has a solution if and only if*

$$L_{\delta}(x) = \min_{\lambda_i \geq 0} \left\| \sum_{i \in \mathcal{J}_{\bar{\delta}}(x)} \lambda_i f'_i(x) \right\| > 0$$

where the minimum is taken over all $\lambda_i \geq 0$ such that

$$\sum_{i \in \mathcal{J}_{\bar{\delta}}(x)} \lambda_i = 1.$$

Then the solution $\bar{p}(x)$ of system (6.20) with a minimum norm satisfies the equality

$$\|\bar{p}(x)\| = \frac{1}{L_{\delta}(x)} F(x).$$

Proof. Let $\lambda_i \geq 0$ be such that their sum over all $i \in \mathcal{J}_{\bar{\delta}}(x)$ is equal to unity. If p is a solution of (6.20), then

$$-\sum_{i \in \mathcal{J}_{\bar{\delta}}(x)} \lambda_i (f'_i(x), p) \geq F(x),$$

or

$$\left(-\sum_{i \in \mathcal{J}_{\bar{\delta}}(x)} \lambda_i f'_i(x), p \right) \geq F(x).$$

Using the inequality $(x, y) \leq \|x\| \|y\|$, we obtain

$$\left\| \sum_{i \in \mathcal{J}_{\bar{\delta}}(x)} \lambda_i f'_i(x) \right\| \|p\| \geq F(x).$$

But the last inequality holds with any λ_i chosen as mentioned above and therefore

$$L_{\delta}(x) \|p\| \geq F(x),$$

i.e. $L_{\delta}(x) > 0$ and

$$\|p\| \geq \frac{F(x)}{L_{\delta}(x)}. \quad (6.21)$$

Thus, it has been proved that the conditions of the lemma are necessary.

Suppose now that $L_{\delta}(x) > 0$

Consider the problem: to find the minimum of ρ with the following constraints:

$$\begin{aligned} (f'_i(x), p) + F(x) - \rho &\leq 0, \quad i \in \mathcal{J}_{\bar{\delta}}(x), \\ \|p\| &\leq r_0, \quad r_0 = \frac{F(x)}{L_{\delta}(x)} > 0. \end{aligned} \quad (6.22)$$

This is a problem of convex programming, and all the conditions of the Kuhn-Tucker theorem, in particular Slater's condition, are

obviously fulfilled. Let p_0, ρ_0 be the solution. Applying the Kuhn-Tucker theorem, we obtain that there are $\lambda_i \geq 0$ such that for all $p, \|p\| \leq r_0$ and for all ρ the following inequality holds:

$$\rho_0 + \sum_{i \in \mathcal{J}_\delta^-(x)} \lambda_i ((f'_i(x), p_0) + F(x) - \rho_0) \leq \rho + \sum_{i \in \mathcal{J}_\delta^-(x)} \lambda_i ((f'_i(x), p) + F(x) - \rho). \quad (6.23)$$

Besides, we have

$$\lambda_i ((f'_i(x), p_0) + F(x) - \rho_0) = 0, \quad i \in \mathcal{J}_\delta^-(x). \quad (6.24)$$

Since ρ is arbitrary, it follows directly from (6.23) that

$$\sum_{i \in \mathcal{J}_\delta^-(x)} \lambda_i = 1.$$

Because of (6.24), we can rewrite (6.23) in the form

$$\rho_0 \leq \left(\sum_{i \in \mathcal{J}_\delta^-(x)} \lambda_i f'_i(x), p \right) + F(x).$$

Taking the minimum with respect to $p, \|p\| \leq r_0$, of the right-hand side of the last inequality we obtain

$$\rho_0 \leq -r_0 \left\| \sum_{i \in \mathcal{J}_\delta^-(x)} \lambda_i f'_i(x) \right\| + F(x) \leq -r_0 L_\delta(x) + F(x) = 0.$$

Thus, $\rho_0 \leq 0$, i.e. vector p_0 satisfies the system of inequalities (see (6.22))

$$(f'_i(x), p_0) + F(x) \leq \rho_0 \leq 0, \quad i \in \mathcal{J}_\delta^-(x),$$

and we have also $\|p_0\| \leq r_0 = F(x)/L_\delta(x)$. But it follows from (6.21) that

$$\|p_0\| \geq \frac{F(x)}{L_\delta(x)}.$$

Therefore

$$\|p_0\| = \frac{F(x)}{L_\delta(x)}$$

and vector p_0 is the solution of system (6.20). Moreover, (6.21) shows that this is a solution with a minimum norm. Thus, $p_0 = \bar{p}(x)$ and the lemma is proved.

Theorem 6.2. *Let all the assumptions of the subsection on p. 211 hold, except the main one. Moreover, let $L_\delta(x) \geq \gamma > 0$ for all x such that $0 < F(x) \leq F(x_k)$. Then the conditions of the main assumption are fulfilled too and all the results of theorem 6.1 hold for problem (6.18).*

Proof. Since any solution of system (6.20) is also the solution of system (6.19), we have

$$\|p(x)\| \leq \|\bar{p}(x)\|.$$

Therefore by lemma 6.1, we have

$$\|p(x)\| \leq \frac{F(x)}{L_\delta(x)} \leq \frac{1}{\gamma} F(x);$$

this shows that all the conditions of theorem 6.1 are satisfied.

Note that the condition $L_\delta(x) \geq \gamma > 0$ is natural enough, for it requires linear independence of vectors $f'_i(x)$, $i \in \mathcal{J}_\delta(x)$.

Theorem 6.3. *Let functions $f_i(x)$ in problem (6.18) be convex and continuously differentiable. Besides, let the domain defined by the inequality $F(x) \leq F(x_0)$ be compact, the gradients $f'_i(x)$ in this domain satisfy Lipschitz' condition and there be a point $\bar{x}$ such that $F(\bar{x}) = \gamma < 0$. Then with $\delta < -\gamma$ all the conditions of theorem 6.1 are fulfilled.*

Proof. As $f_i(x)$ are convex, we have

$$f_i(\bar{x}) \geq f_i(x) + (f'_i(x), \bar{x} - x), \quad i \in \mathcal{J}^-.$$

For $i \in \mathcal{J}_\delta(x)$ with $p = \bar{x} - x$ we have

$$f_i(\bar{x}) + \delta \geq f_i(x) + \delta + (f'_i(x), p).$$

But $f_i(\bar{x}) + \delta \leq F(\bar{x}) + \delta = \gamma + \delta < 0$ and $f_i(x) + \delta \geq F(x)$, $i \in \mathcal{J}_\delta(x)$. Therefore

$$0 \geq \gamma + \delta \geq F(x) + (f'_i(x), p), \quad i \in \mathcal{J}_\delta(x).$$

Setting $\gamma + \delta = -\varepsilon$, we obtain that

$$(f'_i(x), p) + (F(x) + \varepsilon) \leq 0, \quad i \in \mathcal{J}_\delta(x).$$

But by lemma 6.1, this means that for all x such that $F(x) \leq F(x_0)$, $F(x) + \varepsilon \geq 0$, system (6.20) is solvable and for such x also $L_\delta(x) > 0$. Now since the domain $F(x) \leq F(x_0)$ is compact and the functions are continuous, it can be easily ascertained that $L_\delta(x) \geq \gamma > 0$ for all x such that $0 \leq F(x) \leq F(x_0)$.

Thus all the conditions of theorem 6.2 are satisfied and this completes the proof of theorem 6.3.

Solving the Problem of Finding the Minimax

Given functions $f_i(x)$, $i = 1, \dots, m$. We compose the function

$$F(x) = \max_{1 \leq i \leq m} f_i(x). \quad (6.25)$$

The problem is now to find point $x \in E^n$ which minimizes $F(x)$.

It is easy to see that this problem can be reduced to the following one by introducing an additional variable x^{n+1} : to minimize $f_0(x, x^{n+1}) = x^{n+1}$ with constraints

$$f_i(x) - x^{n+1} \leq 0, \quad i = 1, \dots, m.$$

Therefore the methods described above, in particular the linearization method, are now applicable. Note also that in this way we can solve also the problem of the minimization of $F(x)$ if x varies in a certain domain Ω defined by a system of equalities or inequalities.

In this subsection, we shall discuss the method of minimization of $F(x)$ with $x \in E^n$. This method is based on a slight modification of the linearization method.

Let us introduce at each point x the following subsidiary problem:

$$\begin{aligned} \min \left(\beta + \frac{1}{2} \|p\|^2 \right), \\ (f'_i(x), p) + f_i(x) - \beta \leq 0, \quad i \in \mathcal{J}_\delta(x), \end{aligned} \quad (6.26)$$

where $\delta > 0$ and

$$\mathcal{J}_\delta(x) = \{i: 1 \leq i \leq m, \quad f_i(x) \geq F(x) - \delta\}.$$

Note that problem (6.26) is a problem of convex programming for which Slater's condition is satisfied, for taking β sufficiently great we can always satisfy strictly constraints (6.26). By applying directly the Kuhn-Tucker theorem in its differential form, we now find that $p(x)$ and $\beta(x)$ is the solution of problem (6.26) if and only if there are $u^i \geq 0$, $i \in \mathcal{J}_\delta(x)$ such that

$$\begin{aligned} \sum_{i \in \mathcal{J}_\delta(x)} u^i &= 1, \\ p(x) + \sum_{i \in \mathcal{J}_\delta(x)} u^i f'_i(x) &= 0, \\ u^i ((f'_i(x), p(x)) + f_i(x) - \beta(x)) &= 0, \quad i \in \mathcal{J}_\delta(x). \end{aligned} \quad (6.27)$$

Further, point $p = 0$, $\beta = F(x)$, obviously satisfies constraints (6.26). Therefore

$$\beta(x) + \frac{1}{2} \|p(x)\|^2 \leq F(x). \quad (6.28)$$

We formulate now the algorithm for solving the problem. Let x_0 be a certain initial approximation. Let points x_j , $j = 0, 1, \dots, k$ be already constructed. Then

$$x_{k+1} = x_k + \alpha_k p_k \quad (6.29)$$

where $p_k \equiv p(x_k)$. We choose α_k equal to 2^{-i_0} where i_0 is the first of indices $i = 0, 1, \dots$, which satisfies the following inequality:

$$F(x_k + 2^{-i} p_k) \leq F(x_k) - 2^i \varepsilon \|p_k\|^2, \quad \frac{1}{2} < \varepsilon < 1.$$

Thus the condition

$$F(x_{k+1}) \leq F(x_k) - \alpha_k \varepsilon \|p_k\|^2 \quad (6.30)$$

is fulfilled. We formulate now the conditions for convergence of the algorithm.

Lemma 6.2. *$p(x) = 0$ if and only if the necessary conditions for a minimum of $F(x)$ are satisfied at point x .*

To prove the lemma we must recall the necessary conditions for a minimum of $F(x)$ and use an argument analogous to that used for proving lemma 5.1.

Theorem 6.4. *Let $f_i(x)$ be continuously differentiable, domain $\Omega = \{x: F(x) \leq F(x_0)\}$ be bounded and $f'_i(x)$ satisfy in Ω Lipschitz' condition with constant L . Then any limit point x_* of sequence $\{x_k\}$, $k = 0, 1, \dots$, satisfies the necessary conditions for a minimum of $F(x)$ with $x \in E^n$. If $f_i(x)$ are convex, then x_* is the solution of the problem.*

Proof. As in proving theorem 5.1, it is easy to obtain the following estimates:

$$f_i(x_k + \alpha p_k) \leq f_i(x_k) + \alpha (f'_i(x_k), p_k) + \alpha^2 L \|p_k\|^2, \\ i \in \mathcal{J}_\delta(x_k),$$

$$f_i(x_k + \alpha p_k) \leq F(x_k) - \delta + \alpha K \|p_k\|, \quad i \in \mathcal{J}_\delta(x_k)$$

where $K = \max_{i, x \in \Omega} \|f'_i(x)\|$.

If we use now the condition (see (6.26))

$$(f'_i(x_k), p_k) \leq \beta_k - f_i(x_k), \quad \beta_k = \beta(x_k)$$

and also (6.28), then the first estimate takes the following form:

$$f_i(x_k + \alpha p_k) \leq (1 - \alpha) f_i(x_k) + \beta_k \alpha + \alpha^2 L \|p_k\|^2, \\ \leq F(x_k) - \alpha (F(x_k) - \beta_k) + \alpha^2 L \|p_k\|^2 \\ \leq F(x_k) - \frac{\alpha}{2} \|p_k\|^2 + \alpha^2 L \|p_k\|^2.$$

Further, since for $0 \leq \alpha \leq \alpha_k^1$, $\alpha_k^1 = \frac{\delta}{\|p_k\| \left(K + \frac{1}{2} \|p_k\|\right)}$,

$$F(x_k) - \frac{\alpha}{2} \|p_k\|^2 \geq F(x_k) - \delta + \alpha K \|p_k\|,$$

we have

$$f_i(x_k + \alpha p_k) \leq F(x_k) - \frac{\alpha}{2} \|p_k\|^2 + \alpha^2 L \|p_k\|^2 \quad (6.31)$$

with $0 \leq \alpha \leq \alpha_k^1$. Therefore

$$F(x_k + \alpha p_k) \leq F(x_k) - \alpha \|p_k\|^2 \left(\frac{1}{2} - \alpha L\right), \quad 0 \leq \alpha \leq \alpha_k^1. \quad (6.32)$$

If now

$$0 \leq \alpha \leq \alpha_k, \quad \bar{\alpha}_k = \min \left[1, \quad \alpha_k^1, \quad \frac{\varepsilon - \frac{1}{2}}{L} \right], \quad (6.33)$$

then

$$F(x_k + \alpha p_k) \leq F(x_k) - \alpha \|p_k\|^2 \varepsilon.$$

It follows immediately that inequality (6.30) holds if

$$\alpha_k \geq \frac{1}{2} \bar{\alpha}_k \quad (6.34)$$

after a finite number of reductions starting with unity.

It follows immediately from (6.30) that $\alpha_k \|p_k\|^2 \rightarrow 0$. This means that $\|p_k\| \rightarrow 0$. Indeed, $\alpha_k^1 \geq \gamma > 0$ since by (6.27) $\|p(x)\|$ has an upper bound in Ω . But it follows from (6.33), (6.34) that α_k also has a lower bound, a certain positive constant.

Thus, $p_k \rightarrow 0$. Let now x_* be a limit point of the sequence. Without loss of generality, we can take that $x_k \rightarrow x_*$. Moreover, since $u_k^i, i \in \mathcal{J}_\delta(x_k)$ are positive and their sum is equal to unity, we can, setting $u_k^i = 0, i \notin \mathcal{J}_\delta(x_k)$, take that $u_k^i \rightarrow \bar{u}^i$ and $\bar{u}^i \geq 0$, their sum being

$$\sum_{i=1}^m \bar{u}^i = 1. \quad (6.35)$$

We rewrite now (6.27) and (6.26) for points x_k as follows

$$\begin{aligned} p_k + \sum_{i=1}^m u_k^i f'_i(x_k) &= 0, \\ u_k^i ((f'_i(x_k), p_k) + f_i(x_k) - \beta_k) &= 0, \quad i = 1, \dots, m, \\ (f'_i(x_k), p_k) + f_i(x_k) &\leq \beta_k, \quad i \in \mathcal{J}_\delta(x_k). \end{aligned} \quad (6.36)$$

It follows from the last inequality (6.36), if we choose $i \in \mathcal{J}_\delta(x_k)$ such that $f_i(x_k) = F(x_k)$, that

$$\beta_k \geq f_i(x_k) - K \|p_k\| = F(x_k) - K \|p_k\|.$$

But (6.28) shows that $\beta_k \leq F(x_k) - \frac{1}{2} \|p_k\|^2$. Therefore $\beta_k \rightarrow F(x_*)$. Taking the limit in (6.36) we obtain:

$$\begin{aligned} \sum_{i=1}^m \bar{u}^i f'_i(x_*) &= 0, \\ \bar{u}^i (f_i(x_*) - F(x_*)) &= 0, \quad i = 1, \dots, m, \\ \sum_{i=1}^m \bar{u}^i &= 1, \quad \bar{u}^i \geq 0. \end{aligned} \quad (6.37)$$

But these are just the necessary conditions for $F(x)$ to attain its minimum at point x_* (see Chap. I). If $f_i(x)$ are convex, then these conditions will be at the same time sufficient and this proves the theorem.

Let us now give a local estimate of the convergence of the algorithm.

Theorem 6.5. *Let x_* be the minimum point of $F(x)$ and functions $f_i(x)$ be twice continuously differentiable. Besides, let the gradients $f'_i(x_*)$, $i \in \mathcal{J}_0(x_*)$, where $\mathcal{J}_0(x_*) = \{i: f_i(x_*) = F(x_*)\}$, be such that the differences*

$$f'_i(x_*) - f'_{i_0}(x_*), \quad i \neq i_0, \quad i_0 \in \mathcal{J}_0(x_*)$$

are linearly independent and the multipliers $\bar{u}^i$ strictly greater than zero for $i \in \mathcal{J}_0(x_*)$ and $(y, L''(x_*, \bar{u})y) > 0$ for all $y \neq 0$. Here

$$L(x, u) = \sum_{i=1}^m u^i f_i(x) \quad \text{and} \quad L''(x, u)$$

is the matrix of second derivatives with respect to x . Then with sufficiently small $\delta > 0$, and $\alpha > 0$ there is a neighbourhood of point x_* such that the process

$$x_{k+1} = x_k + \alpha p(x_k), \quad k = 0, 1, \dots,$$

converges starting from any initial approximation x_0 of this region and $\|x_* - x_k\| \leq Cq^k$, where $0 < q < 1$.

Proof. We shall give only the general scheme of the proof since the complete proof is quite analogous to the proof of theorem 5.4 and in fact is reduced to it.

If we take

$$\tilde{\mathcal{J}}(x) = \{i \in \mathcal{J}_\delta(x): (f'_i(x), p(x)) + f_i(x) - \beta(x) = 0\},$$

then it can be shown (see lemma 5.2) that with a small δ we have $\mathcal{J}_\delta(x) = \mathcal{J}_0(x_*) = \tilde{\mathcal{J}}(x)$ for all x close to x_* . Therefore it follows from (6.26) and (6.27) that vector $p(x)$ and the corresponding Lagrange multipliers u^i satisfy the system of equations

$$\begin{aligned} p(x) + \sum_{i \in \mathcal{J}_0(x_*)} u^i f'_i(x) &= 0, \\ (f'_i(x), p(x)) + f_i(x) &= \beta(x), \quad i \in \mathcal{J}_0(x_*), \\ \sum_{i \in \mathcal{J}_0(x_*)} u^i &= 1. \end{aligned} \tag{6.38}$$

Let i_0 be an index from $\mathcal{J}_0(x_*)$ and

$$\tilde{f}_i(x) = f_i(x) - f_{i_0}(x), \quad \tilde{f}_0(x) = f_{i_0}(x).$$

Then system (6.38) is equivalent to the following one:

$$\begin{aligned} p(x) + \tilde{f}'_0(x) + \sum_{i \in \tilde{\mathcal{J}}_0(x_*)} u^i \tilde{f}'_i &= 0, \\ (\tilde{f}'_i(x), p(x)) + \tilde{f}_i(x) &= 0, \quad i \in \tilde{\mathcal{J}}_0(x_*), \\ \tilde{\mathcal{J}}_0(x_*) &= \mathcal{J}_0(x_*) \setminus \{i_0\}. \end{aligned} \quad (6.39)$$

But this system is absolutely equivalent to system (5.51.1), (5.51.2). As the proof of theorem 5.4 was reduced to a study of the properties of $p(x)$ —the solution of system (5.51.1), (5.51.2)—it follows that the further proof of theorem 6.5 is simply reduced to checking the conditions of theorem 5.4. But it can be easily ascertained that the assumptions of theorem 6.5 provide completely for the fulfilment of the conditions of theorem 5.4 for functions $\tilde{f}_i$ and this completes the proof of the theorem.

The following theorem is an absolute analogue of theorem 5.5.

Theorem 6.6. *Let the conditions of theorem 6.5 be fulfilled and, besides, the number of indices in set $\mathcal{J}_0(x_*)$ be equal to $n + 1$. In this case, with a small δ the process*

$$x_{k+1} = x_k + p(x_k) \quad (6.40)$$

converges at a quadratic rate to point x_* .

Proof. In the case under consideration, vector $p(x)$ is uniquely defined by the system of equations

$$(\tilde{f}'_i(x), p(x)) + \tilde{f}_i(x) = 0, \quad i \in \tilde{\mathcal{J}}_0(x_*),$$

since vectors $\tilde{f}'_i(x)$, $i \in \tilde{\mathcal{J}}_0(x_*)$ are linearly independent for x , close to x_* by the assumptions. But then process (6.40) is just Newton's method for solving the system of equations

$$\tilde{f}_i(x) = 0, \quad i \in \tilde{\mathcal{J}}_0(x_*) \quad (6.41)$$

which by theorem 6.1 and remark 1 on p. 215, converges quadratically in the neighbourhood of point x_* . Note that point x_* satisfies (6.41), for $f_i(x_*) = F(x_*)$, $i \in \mathcal{J}_0(x_*)$ and therefore

$$\tilde{f}_i(x_*) = f_i(x_*) - f_{i_0}(x_*) = 0, \quad i \in \tilde{\mathcal{J}}_0(x_*).$$

7. LOCAL ACCELERATION OF CONVERGENCE

As was shown in Sec. 5, the linearization method, speaking generally, converges at the rate of a geometric progression. In a number of problems this can prove insufficient and the problem arises of how to accelerate the convergence of the process.

In this section we shall describe methods that permit to do it provided an approximation sufficiently close to the solution has been found. The last circumstance is a shortcoming of the process; there are, however, no methods at present that permit to construct a process, whatever the initial approximation, with an asymptotically superlinear rate of convergence, this being achieved but in the problem of unconstrained minimization.

The methods described below are based on the following idea. The minimization problem is reduced to a certain system of nonlinear equations and then Newton's method or its modification is applied to the solving of this system. At the end of this section we shall describe a method that uses this idea directly, i.e. the necessary conditions for a minimum will be established and Newton's method applied to the solving of the equations obtained. Such a method has many shortcomings of which the principal one is the necessity of calculating second derivatives of the original functions. Therefore, this method can be applied only to problems in which such derivatives are easily calculated.

A second method is based on the fact that point x_* is the solution of minimization problem (5.1) only if it satisfies the equation $p(x_*) = 0$, where vector $p(x)$ is the solution of the subsidiary problem (5.4). We shall describe a method that permits to solve a system of nonlinear equations without calculating derivatives. As was mentioned above, this method will converge only from a sufficiently good initial approximation.

Formulation of the Problem. Basic Formulas

It is required to solve a system of equations

$$p(x) = 0 \quad (7.1)$$

where $p(x)$ is a vector with components $p^i(x)$, $i = 1, \dots, n$, $x \in E^n$. Note that $p(x)$ is an arbitrary vector-function that is not as yet connected with the problem of mathematical programming.

Let x_* be the solution of system (7.1). We shall always assume further on that $p(x)$ is a vector-function which is differentiable in the neighbourhood of point x_* and that the matrix of derivatives

$$p'(x) = \left\{ \frac{\partial p^i(x)}{\partial x^j} \right\}_{\substack{i=1, \dots, n \\ j=1, \dots, n}}$$

satisfies Lipschitz' condition, i.e.

$$\|p'(x) - p'(y)\| \leq L \|x - y\|$$

where all the norms are Euclidean.

Further without loss of generality, we can take that $x_* = 0$. We denote

$$\begin{aligned} p'(0) &= A, \quad \omega(x) = p(x) - Ax, \\ \omega(x, y) &= \frac{1}{\|x - y\|} [p(x) - p(y) - A(x - y)]. \end{aligned} \quad (7.2)$$

Suppose that matrix A is nonsingular so that the following estimates hold:

$$m \|x\| \leq \|Ax\| \leq M \|x\| \quad (7.3)$$

where $M \geq m > 0$.

Lemma 7.1. *The estimates*

$$\|\omega(x)\| \leq C_1 \|x\|^2, \quad \|\omega(x, y)\| \leq C_2 \max \{\|x\|, \|y\|\}$$

hold.

Proof. Let $p^{i'}(x)$ be the gradient of function $p^i(x)$. Then by Taylor's formula we have

$$p^i(x) = p^i(0) + (p^{i'}(0), x) + (p^{i'}(z) - p^{i'}(0), x)$$

where $z = \theta x$, $0 \leq \theta \leq 1$. Using the fact that $p^i(0) = 0$ and Lipschitz' condition for $p'(x)$, we obtain that

$$\|p^i(x) - (p^{i'}(0), x)\| \leq L \|x\|^2, \quad i = 1, \dots, n.$$

Hence, we have

$$\|\omega(x)\| = \|p(x) - p'(0)x\| \leq C_1 \|x\|^2.$$

Further

$$p^i(y) = p^i(x) + (p^{i'}(x), y - x) + (p^{i'}(z) - p^{i'}(x), y - x)$$

where $z = \theta x + (1 - \theta)y$, $0 \leq \theta \leq 1$. Therefore

$$\begin{aligned} p^i(y) - p^i(x) - (p^{i'}(0), y - x) \\ = (p^{i'}(x) - p^{i'}(0), y - x) + (p^{i'}(z) - p^{i'}(x), y - x). \end{aligned}$$

Hence after simple transformations (using Lipschitz' condition), we obtain

$$\begin{aligned} |p^i(y) - p^i(x) - (p^{i'}(0), y - x)| \\ \leq L \|x\| \|y - x\| + L \|z - x\| \|y - x\| \\ = L \|y - x\| (\|x\| + (1 - \theta) \|y - x\|) \\ \leq L \|y - x\| ((2 - \theta) \|x\| + (1 - \theta) \|y\|) \\ \leq 3L \|y - x\| \max \{\|x\|, \|y\|\}. \end{aligned}$$

The second statement of the lemma follows immediately from the last inequality.

Let now points $x_1, x_2, \dots, x_n$ be already constructed, $p(x_k) \neq 0$, $k = 1, \dots, n$, e_k be unit vectors in the direction of the k -th coordinate axis.

We introduce the notations:

$$\begin{aligned} y_k &= x_k + \|p(x_k)\| e_k, \\ r_k &= y_k - x_k = \|p(x_k)\| e_k, \\ z_k &= p(y_k) - p(x_k), \quad k = 1, \dots, n. \end{aligned}$$

We shall introduce a measure of the linear independence of an arbitrary set of vectors b_k , $k = 1, \dots, n$. We take

$$\Delta(b_1, \dots, b_n) = \min_{\sum_{i=1}^n |\alpha_i| = 1} \left\| \sum_{i=1}^n \alpha_i \frac{b_i}{\|b_i\|} \right\|.$$

It is easily seen that $\Delta(b_1, \dots, b_n) > 0$ if and only if vectors $b_1, \dots, b_n$ are linearly independent. Note also that

$$\Delta(e_1, \dots, e_n) = \frac{1}{\sqrt{n}}.$$

Lemma 7.2. *There is a neighbourhood of point $x_* = 0$ such that*

$$\Delta(z_1, \dots, z_n) \geq \gamma > 0,$$

provided $x_1, \dots, x_n$ are in this region.

Proof. By the definition of $\omega(x, y)$, we have

$$z_k = Ar_k + \omega(y_k, x_k) \|r_k\| = \|p(x_k)\| (Ae_k + \omega(y_k, x_k)).$$

Therefore

$$\frac{z_k}{\|z_k\|} = \frac{Ae_k + \omega(y_k, x_k)}{\|Ae_k + \omega(y_k, x_k)\|}.$$

If $x_k \rightarrow 0$, then

$$\frac{z_k}{\|z_k\|} \rightarrow \frac{Ae_k}{\|Ae_k\|}.$$

However, it is easy to see that $\Delta(z_1, \dots, z_n)$ depends continuously on $z_k \|z_k\|^{-1}$. Therefore for x_k sufficiently close to zero we have

$$\Delta(z_1, \dots, z_n) \geq \frac{1}{2} \Delta(Ae_1, \dots, Ae_n).$$

But $\Delta(Ae_1, \dots, Ae_n) > 0$, for vectors Ae_k , $k = 1, \dots, n$ are simply columns of matrix A , and since matrix A is nonsingular, its columns are linearly independent. Thus

$$\Delta(z_1, \dots, z_n) \geq \frac{1}{2} \Delta(Ae_1, \dots, Ae_n) > 0$$

for all x_k from the neighbourhood of $x_* = 0$. Q.E.D.

Let $\delta > 0$ be the radius of the region about zero in which lemmas 7.1 and 7.2 hold. Let points $x_1, \dots, x_n$ be chosen in this region.

Let us find quantities β_i , $i = 1, \dots, n$ from the system of equations

$$-p(x_n) = \sum_{k=1}^n \beta_k z_k. \quad (7.4)$$

By lemma 7.2, this system is solvable. We take

$$x_{n+1} = x_n + \sum_{k=1}^n \beta_k r_k. \quad (7.5)$$

Let us estimate the norm of x_{n+1} . Since

$$\begin{aligned} p(x_n) &= Ax_n + \omega(x_n) \\ z_k &= Ar_k + \omega(y_k, x_k) \|r_k\|, \end{aligned} \quad (7.6)$$

we obtain from (7.4) that

$$-Ax_n - \omega(x_n) = \sum_{k=1}^n \beta_k Ar_k + \sum_{k=1}^n \beta_k \omega(y_k, x_k) \|r_k\|$$

or

$$Ax_{n+1} = -\omega(x_n) - \sum_{k=1}^n \beta_k \omega(y_k, x_k) \|r_k\|.$$

It follows from the last equality that

$$m \|x_{n+1}\| \leq \|Ax_{n+1}\| \leq \|\omega(x_n)\| + \sum_{k=1}^n |\beta_k| \|r_k\| \|\omega(y_k, x_k)\|. \quad (7.7)$$

But by lemma 7.1

$$\begin{aligned} \|\omega(y_k, x_k)\| &\leq C_2 \max \{\|y_k\|, \|x_k\|\} \\ &= C_2 \max \{\|(x_k + \|p(x_k)\| e_k)\|, \|x_k\|\} \\ &\leq C_2 (\|x_k\| + \|p(x_k)\|). \end{aligned}$$

In the region under consideration

$$\|p(x_k)\| = \|Ax_k + \omega(x_k)\| \leq M(x_k) + C_1 \|x_k\|^2.$$

Therefore

$$\|\omega(y_k, x_k)\| \leq C_2 (1 + M + C_1 \delta) \|x_k\| \equiv C_3 \|x_k\|.$$

Using this inequality we can rewrite estimate (7.7) in the following form:

$$\|x_{n+1}\| \leq \frac{1}{m} \left[C_1 \|x_n\|^2 + C_3 \max_{1 \leq k \leq n} \|x_k\| \sum_{k=1}^n |\beta_k| \|r_k\| \right]. \quad (7.8)$$

Note now that

$$\begin{aligned} \left\| \sum_{k=1}^n \beta_k z_k \right\| &= \left\| \sum_{k=1}^n \frac{\beta_k \|z_k\|}{\sum_{i=1}^n |\beta_i| \|z_i\|} \frac{z_k}{\|z_k\|} \right\| \left(\sum_{i=1}^n |\beta_i| \|z_i\| \right) \\ &\geq \Delta(z_1, \dots, z_n) \left(\sum_{k=1}^n |\beta_k| \|z_k\| \right). \end{aligned}$$

Taking into account (7.4), we obtain

$$\|p(x_n)\| \geq \Delta(z_1, \dots, z_n) \left(\sum_{k=1}^n |\beta_k| \|z_k\| \right). \quad (7.9)$$

Further

$$\begin{aligned} \|z_k\| &= \|Ar_k + \omega(y_k, x_k)\| \|r_k\| \\ &\geq \|Ar_k\| - \|r_k\| \|\omega(y_k, x_k)\| \geq \|r_k\| (m - C_3 \|x_k\|) \\ &\geq \|r_k\| (m - C_3 \max_{1 \leq k \leq n} \|x_k\|). \end{aligned}$$

Therefore

$$\sum_{k=1}^n |\beta_k| \|z_k\| \geq \left(\sum_{k=1}^n |\beta_k| \|r_k\| \right) (m - C_3 \max_{1 \leq k \leq n} \|x_k\|).$$

If

$$\max_{1 \leq k \leq n} \|x_k\| < \frac{m}{C_3}, \quad (7.10)$$

then inequality (7.9) yields now

$$\sum_{k=1}^n |\beta_k| \|r_k\| \leq \frac{\|p(x_n)\|}{\Delta(z_1, \dots, z_n) (m - C_3 \max_{1 \leq k \leq n} \|x_k\|)}. \quad (7.11)$$

Taking into account that

$$\begin{aligned} \|p(x_n)\| &= \|Ax_n + \omega(x_n)\| \leq M \|x_n\| + C_1 \|x_n\|^2 \\ &\leq C_4 \|x_n\| \end{aligned}$$

and also (7.11) and lemma 7.2,¹ we can finally rewrite (7.8) as follows:

$$\|x_{n+1}\| \leq \|x_n\| \frac{1}{m} \left[C_1 \|x_n\| + \frac{C_3 \max_{1 \leq k \leq n} \|x_k\| C_4}{\gamma (m - C_3 \max_{1 \leq k \leq n} \|x_k\|)} \right]. \quad (7.12)$$

We formulate now the result obtained in the form of a lemma.

Lemma 7.3. *If points x_k , $k = 1, \dots, n$ are chosen in a region about point $x_* = 0$ such that the conditions of lemmas 7.1 and 7.2 and inequality (7.10) are satisfied, then estimate (7.12) holds true.*

Algorithm

We formulate now the algorithm for solving the system of equations (7.1).

Choose initial points $x_1, x_2, \dots, x_n$. Let points $x_1, \dots, x_n, \dots, x_k$ be already constructed. Then point x_{k+1} is constructed by the following formula:

$$x_{k+1} = x_k + \sum_{i=1}^n \beta_i^k r_{k-n+i} \quad (7.13)$$

where

$$r_j = \|p(x_j)\| e_{m(j)}, \quad y_j = x_j + r_j, \quad z_j = p(y_j) - p(x_j)$$

and the quantities β_i^k , $i = 1, \dots, n$ are determined from the system of equations

$$-p(x_k) = \sum_{i=1}^n \beta_i^k z_{k-n+i}. \quad (7.14)$$

Index $m(j)$ is calculated by the following rule: if $j = ln + p$, $1 \leq p \leq n - 1$, where l is an integer, then $m(j) = p$; and if $j = ln$, then $m(j) = n$.

Thus vectors $r_1, r_2, \dots, r_k$ are proportional to unit vectors of the coordinate axes which are taken in cyclic order.

It can be seen from the above formulas that the scheme of the algorithm is simple enough. At every step it comprises the calculation of $p(x)$ at points x_k and y_k and the solving of the system of equations (7.14).

Theorem 7.1. *Let $\delta_0 > 0$ be such that for all x satisfying the inequality $\|x\| \leq \delta_0$ the conditions of lemmas 7.1 and 7.2 are fulfilled and moreover, the following two inequalities:*

$$m - C_3 \|x\| \geq \frac{m}{2}, \quad (7.15)$$

$$\frac{\delta_0}{m} \left[C_1 + \frac{2C_3C_4}{m\gamma_1} \right] < 1. \quad (7.16)$$

Let $x_1, \dots, x_n$ be chosen such that $\|x_k\| \leq \delta_0$, $k = 1, \dots, n$. Then the algorithm described above converges to the solution x_* of equations (7.1) at a superlinear rate.

Proof. First, we show that $\|x_k\| \leq \delta_0$ for all points x_k constructed by the algorithm. Indeed, if $x_1, \dots, x_k$ are in the δ_0 -region

about point x_* , then the conditions of lemma 7.3 are satisfied for points $x_{k-n+1}, x_{k-n+2}, \dots, x_k$ and therefore the following inequality analogous to inequality (7.12) holds:

$$\|x_{k+1}\| \leq \|x_k\| \frac{1}{m} \left[C_1 \|x_k\| + \frac{C_3 C_4 \max_{1 \leq i \leq n} \|x_{k-n+i}\|}{\gamma (m - C_3 \max_{1 \leq i \leq n} \|x_{k-n+i}\|)} \right].$$

Hence it follows, taking into account (7.15), that

$$\|x_{k+1}\| \leq \|x_k\| \max_{1 \leq i \leq n} \|x_{k-n+i}\| C_5 \quad (7.17)$$

where

$$C_5 = \frac{1}{m} \left[C_1 + \frac{2C_3 C_4}{\gamma m} \right].$$

But $\|x_{k-n+i}\| \leq \delta_0$ by assumption. Therefore

$$\|x_{k+1}\| \leq \|x_k\| \delta_0 C_5 \leq \|x_k\| \leq \delta_0.$$

Q.E.D. Moreover, it follows from the last inequality, with the notation $q_0 = \delta_0 C_5$, that

$$\|x_{k+1}\| \leq q_0 \|x_k\|. \quad (7.18)$$

Since by (7.16) $q_0 < 1$, we have the estimate $\|x_k\| \leq q_0^{k-n} \|x_n\|$, i.e. $x_k \rightarrow 0$, and this proves the first statement of the theorem.

Further, from (7.17) we obtain that

$$\frac{\|x_{k+1}\|}{\|x_k\|} \leq C_5 \max_{1 \leq i \leq n} \|x_{k-n+i}\|. \quad (7.19)$$

Since $x_k \rightarrow 0$, the estimate means that

$$\frac{\|x_{k+1}\|}{\|x_k\|} \rightarrow 0.$$

The last relation shows that $x_k \rightarrow 0$ at a faster rate than that of any geometric progression. The theorem is proved.

We shall give now the more precise bounds on the rate of convergence. We set $v_k = C_5 \|x_k\|$. Then (7.17) can be written in the form

$$v_{k+1} \leq v_k \max_{1 \leq i \leq n} v_{k-n+i}. \quad (7.20)$$

Let us take now $\tilde{v}_j = \max_{1 \leq i \leq n} v_i$, $j = 1, \dots, n$ and determine $\tilde{v}_k$, $k > n$ by the following recursive formula:

$$\tilde{v}_{k+1} = \tilde{v}_k \max_{1 \leq i \leq n} \tilde{v}_{k-n+i}. \quad (7.21)$$

It is now easy to see that $v_k \leq \tilde{v}_k$ for all k . Further, since

$$v_i = C_5 \|x_i\| \leq C_5 \delta_0 = q_0 < 1, \quad i = 1, \dots, n,$$

we have $\tilde{v}_i \leq q_0 < 1$, $i = 1, \dots, n$, and therefore sequence $\{\tilde{v}_k\}$ decreases monotonically. This fact can be proved in an elementary way by induction on k . It follows that $\max_{1 \leq i \leq n} \tilde{v}_{k-n+i} = \tilde{v}_{k-n+1}$ and (7.21) can be rewritten in the form

$$\tilde{v}_{k+1} = \tilde{v}_k \tilde{v}_{k-n+1}. \quad (7.22)$$

We denote $w_k = \ln \tilde{v}_k$. Then

$$\begin{aligned} w_{k+1} &= w_k + w_{k-n+1}, \quad k \geq n, \\ w_k &= \ln \tilde{v}_k, \quad k = 1, \dots, n. \end{aligned} \quad (7.23)$$

It follows from Ostrowski's results (his theorems 12.1 and 12.2) that

$$\frac{w_{k+1}}{w_k} \rightarrow \lambda_0 \quad (7.24)$$

where λ_0 is the greatest positive root of the equation

$$\varphi(\lambda) = \lambda^n - \lambda^{n-1} - 1 = 0. \quad (7.25)$$

As $\varphi(1) = -1 < 0$, and for great λ , $\varphi(\lambda) > 0$, we have $\lambda_0 > 1$.

It follows from (7.24) that for any $\varepsilon > 0$, $\lambda_0 - \varepsilon > 1$ there is a number $k(\varepsilon)$ such that $\frac{w_{k+1}}{w_k} \geq \lambda_0 - \varepsilon$ or $\frac{\ln \tilde{v}_{k+1}}{\ln \tilde{v}_k} \geq \lambda_0 - \varepsilon$. As $\ln \tilde{v}_k < 0$ ($\tilde{v}_k < q_0 < 1$), we have for $k \geq k(\varepsilon)$

$$\ln \tilde{v}_{k+1} \leq (\lambda_0 - \varepsilon) \ln \tilde{v}_k = \ln \tilde{v}_k^{(\lambda_0 - \varepsilon)}$$

or $\tilde{v}_{k+1} \leq \tilde{v}_k^{(\lambda_0 - \varepsilon)}$, $k \geq k(\varepsilon)$. It follows from the last formula that $\tilde{v}_k \leq \tilde{v}_{k(\varepsilon)}^{(\lambda_0 - \varepsilon)(k - k(\varepsilon))}$. But the sequence $\{\tilde{v}_k\}$ decreases monotonically and $\tilde{v}_n \leq q_0 < 1$. Therefore

$$\tilde{v}_k \leq q_0^{(\lambda_0 - \varepsilon)(k - k(\varepsilon))}, \quad k \geq k(\varepsilon). \quad (7.26)$$

Theorem 7.2. *If the conditions of theorem 7.1 are fulfilled, then for every $\varepsilon > 0$, $\lambda_0 - \varepsilon > 1$, where λ_0 is the greatest root of the equation $\lambda^n - \lambda^{n-1} - 1 = 0$, there is a number $k(\varepsilon)$ such that for all $k \geq k(\varepsilon)$ the inequality*

$$\|x_k\| \leq \frac{1}{C_5} q_0^{(\lambda_0 - \varepsilon)(k - k(\varepsilon))}, \quad q_0 < 1 \quad (7.27)$$

holds.

Proof. Recall that $v_k = C_5 \|x_k\|$, $v_k \leq \tilde{v}_k$. These inequalities and (7.26) yield immediately the result required.

Computational Aspects. Application to the Problem of Mathematical Programming

The algorithm described in the preceding subsection is simple enough. It requires at every step the calculation of vector $p(x)$ at points x_k and y_k and the solving of the system of linear equations (7.14). If we denote by Z_k the matrix whose columns are z_{k-n+i} , $i = 1, \dots, n$, then equations (7.14) can be rewritten in the form $Z_k \beta^k = -p(x_k)$, where β^k is a column-vector whose components are β_i^k , $i = 1, \dots, n$.

It follows from the algorithm that matrices Z_k and Z_{k-1} differ only by one column: column z_{k+1} is substituted for column z_k and $z_{k-n+i+1}$, $i \leq n-1$, for z_{k-n+i} , $i \leq n-1$. Therefore to calculate β_i^{k+1} , one can proceed as in the subsections on pp. 76 and 79. Note that the procedure leads to the accumulation of calculation errors. Therefore, if the calculation of $p(y_k)$ requires considerably more operations than the solving of system (7.14), the standard program for solving a system of linear equations should be used for calculating β^k rather than using recursive formulas.

We now turn again to the problem 5.1-5.2 discussed in Sec. 5. According to lemma 5.1, in order to find a local minimum it suffices to solve the equation $p(x) = 0$, where $p(x)$ is the solution of problem (5.4). If the assumptions of theorem 5.4 hold, then by lemmas 5.2, 5.4 in a sufficiently small region about the solution x_* , the conditions of theorem 7.1 are also satisfied. Therefore, the application of the algorithm described in this section makes it possible to accelerate the convergence of the linearization method. In applying this method one should use as $p(x)$ the vector that is the solution of problem 5.4.

Minimization Problem with Equality Constraints

Let us consider the problem of minimization of function $f_0(x)$ with constraints

$$f_i(x) = 0, \quad i = 1, \dots, m. \quad (7.28)$$

Let x_* be the solution of the problem and the following assumptions hold.

(a) Functions $f_i(x)$ are twice continuously differentiable and their second derivatives satisfy Lipschitz' condition.

(b) At point x_* the gradients $f'_i(x_*)$, $i = 1, \dots, m$ are linearly independent so that the necessary conditions for a minimum at x_* are satisfied in their regular form (see Chap. I, Sec. 4). Thus, there are

Lagrange multipliers u^i , $i = 1, \dots, m$ such that

$$\begin{aligned} f'_0(x_*) + \sum_{i=1}^m u^i f'_i(x_*) &= 0, \\ f_i(x_*) &= 0, \quad i = 1, \dots, m. \end{aligned} \quad (7.29)$$

(c) The sufficient conditions for a local minimum, i.e. $(y, L''(x_*, u)y) > 0$, hold if $y \neq 0$ and $(f'_i(x_*), y) = 0$, $i = 1, \dots, m$. Here $L(x, u) = f_0(x) + \sum_{i=1}^m u^i f_i(x)$ and $L''(x, u)$ is the matrix of second derivatives of $L(x, u)$ with respect to x .

Theorem 7.3. *Let the above conditions (a), (b), (c) be fulfilled. Then sequences $\{x_k\}$, $\{u_k^i\}$, $i = 1, \dots, m$, $k = 0, 1, \dots$, calculated by the following recursive formulas*

$$\begin{aligned} L''(x_k, u_k) p_k + \sum_{i=1}^m \Delta u_k^i f'_i(x_k) + L'(x_k, u_k) &= 0, \\ (f'_i(x_k), p_k) + f_i(x_k) &= 0, \quad i = 1, \dots, m, \end{aligned} \quad (7.30)$$

$$\begin{aligned} x_{k+1} &= x_k + p_k, \\ u_{k+1}^i &= u_k^i + \Delta u_k^i, \quad i = 1, \dots, m, \end{aligned} \quad (7.31)$$

converge to x_* and u^i respectively at a quadratic rate, whatever the initial approximation x_0 , u_0^i , $i = 1, \dots, m$ sufficiently close to the solution x_* , u^i , $i = 1, \dots, m$.

Proof. The process defined by formulas (7.30), (7.31) is simply the one generated by Newton's method when it is applied to system (7.29). Therefore, in order to prove the theorem it suffices to check, using remark 1 on p. 215, that the matrix of the first derivatives of the left-hand sides of (7.29) with respect to all x and u^i is nonsingular.

If we denote by $f'(x)$ a matrix whose rows are $f'_i(x)$, $i = 1, \dots, m$, then it is easy to see that the matrix of the first derivatives of the left-hand sides of (7.29) has the form of the following block:

$$\left(\begin{array}{cc} L''(x_*, u) & f'^*(x_*) \\ \underbrace{f'(x_*)}_{n+m} & 0 \end{array} \right) \Bigg\} \begin{matrix} n+m \\ n+m \end{matrix}.$$

In order to ascertain that this matrix is nonsingular, it is sufficient to show that the homogeneous system of equations

$$\begin{aligned} L''(x_*, u) \bar{y} + f'^*(x_*) \bar{u} &= 0, \\ f'(x_*) \bar{y} &= 0 \end{aligned} \quad (7.32)$$

has only a zero solution. In this system, $\bar{y} \in E^n$, $\bar{u}$ is a vector whose components are $\bar{u}^i$, $i = 1, \dots, m$. Let $\bar{y}$, $\bar{u}$ be the solution of system (7.32). Performing scalar multiplication of the first of equations (7.32) by $\bar{y}$, we obtain on the basis of the second equation that

$$\begin{aligned} & (\bar{y}, L''(x_*, u) \bar{y}) + (\bar{y}, f'^*(x_*) \bar{u}) \\ &= (\bar{y}, L''(x_*, u) \bar{y}) + (f'(x_*) \bar{y}, \bar{u}) = (\bar{y}, L''(x_*, u) \bar{y}) = 0. \end{aligned}$$

But by assumption (c), the last expression shows that $\bar{y} = 0$. Therefore, the first of relations (7.32) can be rewritten as follows:

$$f'^*(x_*) \bar{u} \equiv \sum_{i=1}^m \bar{u}^i f'_i(x_*) = 0,$$

and this is possible only if $\bar{u}^i = 0$, $i = 1, \dots, m$ since vectors $f'_i(x_*)$ are linearly independent by assumption (b).

Thus we have shown that the conditions of convergence of Newton's method are fulfilled and consequently the theorem is proved.

8. METHOD OF PENALTY FUNCTIONS

The method of penalty functions is one of the simplest and widely known of the methods for solving the problem of mathematical programming. The basic idea of the method consists in approximately reducing the constrained minimization problem to the unconstrained minimization of a certain function. The subsidiary function is chosen so that it coincides with the function to be minimized in the admissible domain and increases steeply outside it.

Suppose we study the problem of minimization of function $f_0(x)$, $x \in E^n$ with constraints

$$f_i(x) \leq 0, \quad i = 1, \dots, m. \quad (8.1)$$

All the functions $f_i(x)$, $i = 0, 1, \dots, m$ are continuous.

We introduce the notations

$$\varphi_0(t) = \begin{cases} t^2, & t \geq 0, \\ 0, & t < 0; \end{cases} \quad \varphi_1(t) = \begin{cases} t, & t \geq 0, \\ 0, & t < 0. \end{cases} \quad (8.2)$$

Let us compose a function

$$\psi(x, r) = r \sum_{i=1}^m \varphi_0(f_i(x)). \quad (8.3)$$

It is easy to see that

$$\psi(x, r) = 0, \quad x \in \Omega$$

where

$$\Omega = \{x: f_i(x) \leq 0, \quad i = 1, \dots, m\}.$$

If $x \notin \Omega$, then $\psi(x, r) > 0$ and $\psi(x, r) \rightarrow +\infty$ as $r \rightarrow +\infty$. The subsidiary problem is now the minimization of the function

$$F(x, r) \equiv f_0(x) + \psi(x, r). \quad (8.4)$$

It is natural to expect that the solution of this problem $x(r)$ will be close to the solution of the original one. The precise conditions under which this fact will be realized are formulated below.

Note that the choice of function $\psi(x, r)$ in the way it was done above is not the only one permissible. It is sufficient for the function to have certain general properties that ensure the convergence of the method. The method has different properties depending on the way of choosing functions $\psi(x, r)$. In particular, if we set

$$\psi(x, r) = r \max_{1 \leq i \leq r} \varphi_i(f_i(x)),$$

then the linearization method described in Sec. 5 can be taken as suitable for the minimization of function (8.4). In this case, as was shown in Sec. 5, it is not necessary that r should tend to infinity. However, $F(x, r)$ will not be a smooth function.

In the general case, $F(x, r)$ is constructed so as to be smooth and make it possible to apply one of the methods of Chap. II, which converge at a fast rate. Unfortunately in this case, r must tend to infinity and this fact involves a number of implicit difficulties which, in the opinion of the authors, considerably impair the value of the penalty function method. We shall discuss below these difficulties and give a cursory description of one more method whose idea is close to that of the penalty function method.

Substantiation of the Penalty Function Method

Let a certain continuous function $\psi(x, r)$ have the following properties:

- (1) $\psi(x, r) = 0$ if $x \in \Omega$, $\psi(x, r) > 0$, $x \notin \Omega$, and $\psi(x_k, r_k) \rightarrow +\infty$ if $x_k \rightarrow x_0$, $x_0 \notin \Omega$, $r_k \rightarrow +\infty$;
- (2) $\psi(x, r)$ increases monotonically with increasing r .

Theorem 8.1. *Let the set*

$$\Omega_C(r) = \{x: F(x, r) \leq C\},$$

$$F(x, r) = f_0(x) + \psi(x, r)$$

be compact. Then function $F(x, r)$ assumes its minimum $m(r)$ for all x at a certain point $x(r)$ and $m(r) \leq m$, where

$$m = \min_{x \in \Omega} f_0(x), \quad m(r) \rightarrow m$$

and $m(r)$ increases monotonically with increasing r . Moreover, if $x(r_k) \rightarrow x_0$, $k \rightarrow \infty$, $r_k \rightarrow \infty$, then x_0 is the solution of the original problem.

Proof. Let $\bar{x}$ be a point of Ω , $\bar{C} = f_0(\bar{x})$. Then set $\bar{\Omega}$ of those $x \in \Omega$ for which $f_0(x) \leq \bar{C}$ is a closed subset of compact set $\Omega_{\bar{C}}(r)$. Indeed, for $x \in \bar{\Omega}$ by the properties of $\psi(x, r)$ we have

$$f_0(x) + \psi(x, r) = f_0(x) \leq \bar{C},$$

i.e. $x \in \Omega_{\bar{C}}(r)$. But it is clear that the minimum of $f_0(x)$ in Ω must be in subset $\bar{\Omega}$. Therefore, we should seek the minimum of continuous function $f_0(x)$ in compact set $\bar{\Omega}$. Since the continuous function attains its minimum in a compact set, it follows that problem (8.1) is solvable. An analogous reasoning shows that function $F(x, r)$ assumes its minimum $m(r)$ at a certain point $x(r)$.

Let x_* be a certain point of the minimum of $f_0(x)$ in Ω . Then

$$F(x_*, r) = f_0(x_*) + \psi(x_*, r) = f_0(x_*).$$

for $x_* \in \Omega$ and $\psi(x, r) = 0$ with $x \in \Omega$. Therefore

$$\min_x F(x, r) = m(r) \leq m,$$

i.e. $x(r) \in \Omega_m(r)$.

Consider now the sets

$$\Omega_m(r) = \{x: f_0(x) + \psi(x, r) \leq f_0(x_*)\}.$$

These sets are compact by assumption, and because of the increase of $\psi(x, r)$ with increasing r , we have

$$\Omega_m(r_2) \subset \Omega_m(r_1), \quad r_1 < r_2.$$

Let now $\{r_k\}$, $k \rightarrow \infty$ be an increasing sequence of r and $r_k \rightarrow +\infty$. Then,

$$\Omega_m(r_k) \subset \Omega_m(r_1).$$

Since as was shown above $x(r) \in \Omega_m(r)$, all points $x(r_k)$ belong to compact set $\Omega_m(r_1)$. Therefore without loss of generality, we can take that sequence $\{x(r_k)\}$ converges to a certain point x_0 .

Let us show that $x_0 \in \Omega$ and $f_0(x_0) = m$. Indeed, if $x_0 \notin \Omega$, then $\psi(x(r_k), r_k) \rightarrow +\infty$ and consequently $F(x(r_k), r_k) \rightarrow +\infty$, for

$$f_0(x(r_k)) \geq \min_{x \in \Omega_m(r_1)} f_0(x).$$

But this contradicts the fact that $F(x(r_k), r_k) = m(r_k) \leq m$. Thus $x_0 \in \Omega$.

Further,

$$\begin{aligned} m(r_k) = F(x(r_k), r_k) &= f_0(x(r_k)) + \psi(x(r_k), r_k) \\ &\leq f_0(x_0) + \psi(x_0, r_k) = f_0(x_0). \end{aligned}$$

Hence

$$\lim_{k \rightarrow \infty} m(r_k) = \lim_{k \rightarrow \infty} (f_0(x(r_k)) + \psi(x(r_k), r_k)) \leq f_0(x_0). \quad (8.5)$$

But $f_0(x(r_k)) \rightarrow f_0(x_0)$. Therefore

$$\lim_{k \rightarrow \infty} \psi(x(r_k), r_k) \leq f_0(x_0) - \lim_{k \rightarrow \infty} f_0(x(r_k)) = 0.$$

As $\psi(x, r) \geq 0$, it follows that

$$\lim_{k \rightarrow \infty} \psi(x(r_k), r_k) = 0.$$

Thus,

$$\lim_{k \rightarrow \infty} m(r_k) = \lim_{k \rightarrow \infty} f_0(x(r_k)) + \lim_{k \rightarrow \infty} \psi(x(r_k), r_k) = f_0(x_0) \geq m.$$

On the other hand, $m(r_k) \leq m$. Therefore $\lim_{k \rightarrow \infty} m(r_k) = f_0(x_0) \leq m$.

Comparing the last inequality with the preceding one, we see that $\lim_{k \rightarrow \infty} m(r_k) = f_0(x_0) = m$, and this completes the proof of the theorem.

The theorem proved shows that the substitution of the minimization of function $F(x, r)$ for the solving of problem (8.1) with great r permits us to come nearer to the solution of the original problem. Let us estimate the character of this convergence for the nonconvex case. The problem of convex programming will be studied in the next subsection.

Theorem 8.2. *Let functions $f_i(x)$, $i = 0, 1, \dots, m$ be continuously differentiable and the conditions of theorem 8.1 be satisfied. Moreover, we take that:*

- (1) *problem (8.1) has a unique solution,*
- (2) *function $\psi(x, r)$ is chosen in form (8.3) and the minimum of $F(x, r)$ is attained at a unique point $x(r)$ with great r ,*
- (3) *in the solution x_* of problem (8.1) the gradients $f'_i(x_*)$, $i \in \mathcal{J}_0(x_*)$ are linearly independent and*

$$\mathcal{J}_0(x_*) = \{i: f_i(x_*) = 0, i = 1, \dots, m\}.$$

Then $\lim_{r \rightarrow \infty} x(r) = x_*$ and

$$\lim_{r \rightarrow \infty} r\varphi_1(f_i(x(r))) = \frac{1}{2} u^i, \quad i = 1, \dots, m$$

where u^i are Lagrange multipliers of problem (8.1).

Remark. Recall that according to theorem 4.1 (Chap. I) the necessary conditions for a minimum are fulfilled at point x_* in the follow-

ing form:

$$f'_0(x_*) + \sum_{i=1}^m u^i f'_i(x_*) = 0, \\ u^i \geq 0, \quad u^i f_i(x_*) = 0, \quad i = 1, \dots, m. \quad (8.6)$$

Proof. First, we shall show that $x(r) \rightarrow x_*$. Suppose that the opposite holds. Then there is a sequence $r_k \rightarrow \infty$ such that $\|x(r_k) - x_*\| \geq \delta_0 > 0$. Since in proving theorem 8.1 it was shown that $x(r_k) \in \Omega_m(r_1)$ and set $\Omega_m(r_1)$ is compact, we can take, without loss of generality (if required we can take a subsequence), that $x(r_k) \rightarrow x_{**}$ and clearly $\|x_{**} - x_*\| \geq \delta_0 > 0$. However, it follows from theorem 8.1 that x_{**} is the solution of problem (8.1). Thus we have obtained two different solutions of problem (8.1). But this contradicts the assumption.

Thus, it has been shown that $x(r) \rightarrow x_*$. We turn to the proof of the second statement of the theorem. As $x(r)$ is the minimum point of function $F(x, r)$, at this point the gradient of the function

$$F(x, r) = f_0(x) + r \sum_{i=1}^m \varphi_0(f_i(x))$$

must be equal to zero. Simple calculations show that we come to the equality

$$F'(x(r), r) = f'_0(x(r)) + \sum_{i=1}^m (2r\varphi_1(f_i(x(r)))) f'_i(x(r)) = 0$$

or, with the notation

$$u^i(r) = 2r\varphi_1(f_i(x(r))),$$

to the equality

$$f'_0(x(r)) + \sum_{i=1}^m u^i(r) f'_i(x(r)) = 0. \quad (8.7)$$

Note now that as $x(r) \rightarrow x_*$ we have $f_i(x(r)) < 0$ for $i \in \mathcal{J}_0(x_*)$ since $f_i(x_*) < 0$, $i \in \mathcal{J}_0(x_*)$. Therefore with great r

$$u^i(r) = 2r\varphi_1(f_i(x(r))) = 0, \quad i \in \mathcal{J}_0(x_*).$$

But $u^i f_i(x_*) = 0$ and therefore for $i \in \mathcal{J}_0(x_*)$ we have $u^i = 0$. Thus taking into account the expression for $u^i(r)$, the statement of the theorem is proved for $i \in \mathcal{J}_0(x_*)$.

Due to the foregoing, we can rewrite (8.7) and (8.6) as follows:

$$f'_0(x(r)) + \sum_{i \in \mathcal{J}_0(x_*)} u^i(r) f'_i(x(r)) = 0, \\ f'_0(x_*) + \sum_{i \in \mathcal{J}_0(x_*)} u^i f'_i(x_*) = 0. \quad (8.8)$$

If we take into account that $x(r) \rightarrow x_*$, $f'_i(x)$ are continuous for all x and $f'_i(x_*)$, $i \in \mathcal{J}_0(x_*)$ are linearly independent, then it is easy to conclude from (8.8) that $u^i(r) \rightarrow u^i$ and this, with account taken of the expression for $u^i(r)$, completes the proof of the theorem.

Remark. It follows from theorem 8.2 that if $u^i > 0$, then with great r functions $f_i(x(r))$ are strictly greater than zero and $f_i(x(r))$ tends to zero at the same rate as quantity $u^i r^{-1}$ does. Thus the approximate solution will always violate the constraint $f_i(x) \leq 0$ if $u^i > 0$.

Convex Programming

In the case of the problem of convex programming, the estimates of the approximation of $x(r)$ to the sought solution x_* can be made more precise.

Theorem 8.3. *Let all the functions $f_i(x)$, $i = 0, 1, \dots, m$ be convex, the conditions of theorem 8.1 for function $\psi(x, r)$ in form (8.3) be satisfied and, besides, the necessary conditions in the form of the Kuhn-Tucker theorem be fulfilled at point x_* which is the solution of problem 8.1, i.e. there are numbers $u^i \geq 0$ such that*

$$\begin{aligned} f_0(x_*) &\leq \sum_{i=1}^m u^i f_i(x) + f_0(x), \text{ for all } x, \\ u^i f_i(x_*) &= 0, \quad i = 1, \dots, m. \end{aligned} \quad (8.9)$$

Then

$$f_i(x(r)) \leq \frac{\bar{u}}{r}, \quad \text{if } f_i(x(r)) \geq 0, \quad (8.10)$$

$$f_0(x(r)) \geq f_0(x_*) - \frac{5}{4} \frac{\bar{u}^2}{r} \quad (8.11)$$

where

$$\bar{u} = \sqrt{\sum_{i=1}^m (u^i)^2}.$$

Proof. We introduce the notation $\tilde{\mathcal{J}}_0(x) = \{i: f_i(x) \geq 0, i = 1, \dots, m\}$. As

$$F(x(r), r) = f_0(x(r)) + r \sum_{i=1}^m \varphi_0(f_i(x(r))) \leq f_0(x_*),$$

it follows from (8.9) that

$$f_0(x(r)) + r \sum_{i=1}^m \varphi_0(f_i(x(r))) \leq f_0(x(r)) + \sum_{i=1}^m u^i f_i(x(r))$$

or

$$r \sum_{i=1}^m \varphi_0(f_i(x(r))) \leq \sum_{i=1}^m u^i f_i(x(r)).$$

But for $i \notin \tilde{\mathcal{J}}_0(x(r))$

$$\varphi_0(f_i(x(r))) = 0, \quad f_i(x(r)) < 0,$$

and $\varphi_0(f_i(x(r))) = f_i^2(x(r))$ for $i \in \tilde{\mathcal{J}}_0(x(r))$. Therefore the inequality obtained may be made stronger:

$$r \sum_{i \in \tilde{\mathcal{J}}_0(x(r))} f_i^2(x(r)) \leq \sum_{i \in \tilde{\mathcal{J}}_0(x(r))} u^i f_i(x(r)) \leq \bar{u} \sqrt{\sum_{i \in \tilde{\mathcal{J}}_0(x(r))} f_i^2(x(r))}.$$

In deriving the last inequality we used the well known Cauchy-Buniakowski inequality.

Thus

$$f_i(x(r)) \leq \sqrt{\sum_{i \in \tilde{\mathcal{J}}_0(x(r))} f_i^2(x(r))} \leq \frac{\bar{u}}{r} \quad (8.12)$$

and it follows that (8.10) holds.

Further, for all x

$$\begin{aligned} f_0(x_*) &\leq f_0(x) + \sum_{i=1}^m u^i f_i(x) \leq f_0(x) + \sum_{i \in \tilde{\mathcal{J}}_0(x)} u^i f_i(x) \\ &= f_0(x) + r \sum_{i \in \tilde{\mathcal{J}}_0(x)} f_i^2(x) - \sum_{i \in \tilde{\mathcal{J}}_0(x)} \left(\sqrt{r} f_i(x) - \frac{u^i}{2\sqrt{r}} \right)^2 \\ &\quad + \sum_{i \in \tilde{\mathcal{J}}_0(x)} \frac{(u^i)^2}{4r} \leq f_0(x) + r \sum_{i=1}^m \varphi_0(f_i(x)) + \frac{\bar{u}^2}{4r} = F(x, r) + \frac{\bar{u}^2}{4r}. \end{aligned}$$

In writing this formula it was taken into account that

$$\sum_{i=1}^m \varphi_0(f_i(x)) = \sum_{i \in \tilde{\mathcal{J}}_0(x)} f_i^2(x), \quad (8.13)$$

by the definition of $\varphi_0(x)$ and $\tilde{\mathcal{J}}_0(x)$. Thus

$$f_0(x_*) \leq f_0(x(r)) + r \sum_{i=1}^m \varphi_0(f_i(x(r))) + \frac{\bar{u}^2}{4r}.$$

But it follows from (8.12) and (8.13) that

$$\sum_{i=1}^m \varphi_0(f_i(x(r))) \leq \frac{\bar{u}^2}{4r^2}.$$

Therefore

$$f_0(x(r)) \geq f_0(x_*) - \frac{5}{4} \frac{\bar{u}^2}{r}.$$

Q.E.D.

Computational Aspects

The methods expounded above reduce problem (8.1) to the minimization of function $F(x, r)$. It is possible now, in order to obtain an approximate solution, to use one of the methods described in Chap. II. The following specific circumstances should be, however, taken into account. If functions $f_i(x)$ are not convex, then function $F(x, r)$ is also not convex with respect to x . Therefore, it can have local minima while in all the preceding text it was assumed that we were determining the global minimum $x(r)$.

As all the methods of Chap. II are meant for finding of a local minimum, if the function to be minimized is not convex, it is the local minimum that will be found if the initial approximation is poor. This affects the convergence and is an important shortcoming of the penalty function method in its application to nonconvex problems.

If the problem under consideration is one of convex programming with the use of function (8.3) as $\psi(x, r)$, then it is easily ascertained that $F(x, r)$ is convex too; therefore, the difficulty mentioned above is removed. However, another difficulty arises. The fact is that one should take r sufficiently great in order to obtain a good approximation; this follows from the estimates obtained above. In this case all derivatives of $F(x, r)$ with respect to x will also be great, for they are proportional to r . But it was established in analysing all methods described in Chap. II whose rate of convergence is superlinear that the size of the region in which the rate of convergence becomes superlinear is in inverse proportion to Lipschitz' constant of second derivatives, i.e. in the case under consideration this region will also be small and even a method which, at the limit, theoretically converges rapidly can become ineffective. Moreover, as function $\varphi_0(t)$ with $t = 0$ has no second derivative, $F(x, r)$ calculated by formula (8.3) with the use of $\psi(x, r)$ will also have no second derivatives at points x for which $f_i(x) = 0$ for a certain i . But if the solution x_* lies on the boundary of the domain, it is this case that will take place. On the other hand, all methods which converge

at a fast rate require that the function being minimized have second derivatives at least in a certain region about the point sought.

All of the difficulties mentioned are, as a rule, observed in calculations in practice and this lowers the effectiveness of the method.

Flacco and McCormick Method

This method is practically applicable only to problems of convex programming. It is based on an idea close to that of the penalty function method; however, in this method the approximations approach the solution from inside the domain and not from outside as in the penalty function method.

Let us again consider problem (8.1). Suppose that all the functions $f_i(x)$ are convex and there is a point $\bar{x}$ such that $f_i(\bar{x}) < 0$, $i = 1, \dots, m$, so that the interior of the admissible set Ω is non-empty. We compose the function

$$P(x, r) = f_0(x) - \sum_{i=1}^m \frac{r}{f_i(x)}, \quad r > 0$$

defined inside set Ω . It is easily ascertained that $P(x, r)$ is convex with respect to x inside Ω . If we denote by $x(r)$ the minimum point of $P(x, r)$ in Ω , then with sufficiently general assumptions analogous to those of theorems 8.1 and 8.2 it can be shown that

$$\lim_{r \rightarrow +0} x(r) = x_*,$$

$$\lim_{r \rightarrow 0} \frac{r}{f_i^2(x(r))} = u^i, \quad i = 1, \dots, m.$$

Thus, the approximate method of solving problem (8.1) again has been reduced to the problem of unconstrained minimization of function $P(x, r)$.

Of the specific traits of this subsidiary problem the same things can be said that were said of the penalty function method in the subsection on p. 242. In order to illustrate these traits and to show why even effective methods of minimization of $F(x, r)$ or $P(x, r)$ may fail to provide for a fast rate of convergence, we shall adduce a simple example.

Let $f_0(x) = -x$, $f_1(x) = x$, $x \in E^1$, i.e. we are solving the problem of minimization of $-x$ subject to $x \leq 0$. The obvious solution is $x_* = 0$:

$$P(x, r) = -x - \frac{r}{x}.$$

Equating to zero the derivative of $P(x, r)$ with respect to x we obtain

$$P'(x, r) = -1 + \frac{r}{x^2} = 0. \quad (8.14)$$

Hence $x(r) = -\sqrt{r}$. Let us now apply to the solving of (8.14) a method that converges at a quadratic rate—Newton's method, i.e. obtain approximations by the formula

$$x_{k+1} = x_k - \frac{P'(x_k, r)}{P''(x_k, r)}.$$

Substituting expressions for $P'(x, r)$ and $P''(x, r)$ we obtain after simple transformations

$$v_{k+1} = \frac{2\sqrt{r} - x_k}{2r} v_k^2, \quad v_k = x_k + \sqrt{r}. \quad (8.15)$$

It is clear from formula (8.15) that the deviation of x_k from the solution $x(r) = -\sqrt{r}$ tends to zero monotonically only with initial points such that

$$\frac{2\sqrt{r} - x_k}{2r} |v_k| < 1.$$

As $x_k < 0$ (the approximation is sought in the region $x < 0$), we have

$$\frac{2\sqrt{r}}{2r} |v_k| < 1, \quad |v_k| < \sqrt{r}.$$

Thus the last formula shows that a quadratic rate of convergence of Newton's method will be guaranteed only in a domain such that in it x_k deviates from the solution by not more than $\sqrt{r}$, i.e. the domain of convergence of Newton's method tends to zero with decreasing r , and the size of this domain is of the order of magnitude of the deviation of $x(r)$ from the true solution of the problem— x_* . This indicates that the greater amount of calculation work will be required to hit the region of convergence of Newton's method, while in cases where Newton's method has a good convergence it is no more necessary as the approximation obtained deviates from x_* by as much as $x(r)$ does.

9. PROJECTION METHODS WITH RESTORATION OF TIES

Construction of the Methods

Consider the problem of minimization of function $f_0(x)$ with the following conditions:

$$f_i(x) = 0, \quad i = 1, \dots, m, \quad m < n. \quad (9.1)$$

We set $g = (f_1, \dots, f_m)$, $S_g = \{x: g(x) = 0\}$ and suppose that all functions $f_0(x), f_1(x), \dots, f_m(x)$ are continuously differentiable and S_g is a smooth manifold $((n - m)$ dimensional), i.e. that at any point $x \in S_g$ the rank of matrix $g'(x)$ is equal to m ($g'(x) = \left\{ \frac{\partial f_i(x)}{\partial x^j} \right\}$, $i = 1, \dots, m$, $j = 1, \dots, n$, i is the row index).

Consequently, at any point $\tilde{x} \in S_g$ a hyperplane tangent to S_g can be constructed:

$$g'(\tilde{x})(x - \tilde{x}) = 0. \quad (9.2)$$

Further on, we denote this hyperplane (i.e. the set of points that satisfy equation (9.2)) by $T(\tilde{x})$.

One possible approach to the construction of iterative processes for solving the problem formulated is based on the following considerations.

Let x_0 be an arbitrary point of S_g such that the gradient $f'_0(x_0)$ is not orthogonal to the hyperplane $T(x_0)$ (i.e. the necessary condition for an extremum of function $f_0(x)$ on the manifold S_0 is not fulfilled at point x_0). Then in plane $T(x_0)$ there are infinitely many directions of descent of $f_0(x)$ (i.e. there are infinitely many directions $x - x_0$ which belong to $T(x_0)$ and such that $(f'_0(x_0), x - x_0) < 0$). Suppose we have determined one of these directions $v_0 = \bar{x}_0 - x_0$ and constructed point $\hat{x}_0(\alpha) = x_0 + \alpha v_0$ such that $f_0(\hat{x}_0(\alpha)) < f_0(x_0)$. Point $\hat{x}_0$ no longer satisfies equations of conditions (9.1). However, if the value of parameter α is sufficiently small (the quantity $\|x_0 - \hat{x}_0(\alpha)\|$ is small), then using point $\hat{x}_0(\alpha)$ we can construct in several ways point $x_1 \in S_g$ such that

$$f_0(x_1) < f_0(x_0). \quad (9.3)$$

This statement is based on the fact that we can choose a point $x_1(\alpha)$ on the smooth manifold S_g (and not a single one) so that the following condition be fulfilled:

$$x_1(\alpha) - x_0 = \hat{x}_0(\alpha) - x_0 + \omega_0(\alpha)$$

where

$$\|\omega_0(\alpha)\| \equiv \|x_1(\alpha) - \hat{x}_0(\alpha)\| = o(\|\hat{x}_0(\alpha) - x_0\|). \quad (9.4)$$

(This can be proved strictly by using the theorem of mapping on one another of the region about point x_0 in manifold S_g and in the tangent manifold $T(x_0)$; the theorem holds in space E^n ; see L. A. Luster-nik and V. I. Sobolev.)

If (9.4) holds, we have, since $f_0(x)$ is differentiable,

$$\begin{aligned} f_0(x_1) &= f_0(\hat{x}_0) + (f'_0(\hat{x}_0), x_1 - \hat{x}_0) + o(\|x_1 - \hat{x}_0\|) \\ &= f_0(x_0) + (f'_0(x_0), \hat{x}_0 - x_0) + o(\|\hat{x}_0 - x_0\|) \\ &\quad + (f'_0(\hat{x}_0), x_1 - \hat{x}_0) + o(\|x_1 - \hat{x}_0\|) \\ &= f_0(x_0) + (f'_0(x_0), \hat{x}_0 - x_0) + o_1(\|\hat{x}_0 - x_0\|). \end{aligned}$$

Hence if parameter α is sufficiently small, the inequality (9.3) is satisfied.

By constructing point $x_1 \in S_g$ at which condition (9.3) is fulfilled, we have performed in essence an iteration of a certain process of descent for constructing successive approximations to the solution. Thus the k -th iteration of the process of the type being described consists in the following.

1. The direction of descent $v_k = \bar{x}_k - x_k$ of function $f_0(x)$ in the tangent hyperplane $T(x_k)$ is determined.

2. A step of definite length is made in the direction v_k : $\hat{x}_k(\alpha) = x_k + \alpha v_k$ (so that $f_0(\hat{x}_k) < f_0(x_k)$).

3. Using point $\hat{x}_k(\alpha)$, point $x_{k+1} \in S_g$ is determined such that condition $f_0(x_{k+1}) < f_0(x_k)$ be satisfied.

It is clear from the foregoing that we can choose for moving from point x_k different directions of descent in plane $T(x_k)$. The choice of quantity α_k and the final step of the iteration—construction of point x_{k+1} —are also determined not uniquely. Performing in different ways each of the three stages of the iteration we can construct a whole class of processes of descent of the type described.

Consider now several possible methods of choosing vector v_k . We can take as vector v_k the projection of the antigradient $-f'_0(x_k)$ onto the plane $T(x_k)$. The construction of such a vector is equivalent to the solving of the problem of minimization of function

$$F_k(x) = (f'_0(x_k), x - x_k) + \frac{1}{2} \|x - x_k\|^2 \quad (9.5)$$

provided that $x \in T(x_k)$. Applying the method of Lagrange multipliers, we find that

$$v_k = -(I - g'^*(g'g'^*)^{-1}g') f'_0(x_k) \quad (9.6)$$

where $g' = g'(x_k)$.

More effective projection methods with restoration of ties can be constructed by choosing as v_k a vector that minimizes the function

$$F_k(x) = (f'_0(x_k), x - x_k) + \frac{1}{2} (f''_0(x_k)(x - x_k), x - x_k) \quad (9.7)$$

on plane $T(x_k)$ (there is such a vector if F_k is a convex function). Since in this case the quadratic approximation to the function being minimized was practically used for the construction of the direction of motion, we shall call the methods, in which v_k is constructed in the way described, *methods of the second order*.

Consider now the method of restoration of ties (the third stage of an iteration) which will be used in what follows.

Let the system of equations (9.1) in a certain region about any point $x \in S_g$ define the function $y = y(z)$, where y is an m -dimension-

al vector of coordinates and z is an $(n - m)$ -dimensional vector. Without loss of generality, we can take $y = (x^1, \dots, x^m)$, $z = (x^{m+1}, \dots, x^n)$. By the theorem of implicit functions, it is necessary for the existence of function $y(z)$ and its derivatives that at any point $x \in S_g$ we have the determinant

$$|g_y(x)| = \left| \left\{ \frac{\partial f_i}{\partial x^j} \right\} \right| \neq 0, \quad i, j = 1, \dots, m. \quad (9.8)$$

In this case, point $x_{k+1} = (z_{k+1}, y_{k+1})$ can be constructed by the formulas

$$z_{k+1} = z_k + \alpha_k p_k, \quad y_{k+1} = y(z_{k+1}) \quad (9.9)$$

where $p_k = \bar{z}_k - z_k$ is the corresponding part of vector v_k . To construct sequence (9.9), there is no need in finding the explicit expression of function $y(z)$; it is sufficient to be able to evaluate it (i.e. to solve system (9.1)) with a fixed vector z .

The construction of sequence (9.9) (besides the realization in this way of one of the possible methods of restoration of ties) can be considered in another respect—as an iterative process of minimization of function $\varphi(z) = f_0(z, y(z))$. Obviously, with condition (9.8) fulfilled, the minimization of $\varphi(z)$ is equivalent to the solving of the original problem. Vector p_k (which is the direction of descent of function $\varphi(z)$) can be considered in this case to be the solution of the problem of minimization of function $\psi_k(z) = F_k(z, y_l(z))$, where function $F_k(z, y)$ is determined by formula (9.5) or (9.7), and vector-function $y_l(z)$ is determined from the linearized equation of ties (i.e. from the equation of the tangent plane $T(x_k)$)

$$g_y(x_k)(y - y_k) + g_z(x_k)(z - z_k) = 0.$$

Hence

$$y_l(z) = y_k - g_y^{-1}(x_k) g_z(x_k)(z - z_k).$$

The fact that vector p_k determined by the methods described is the direction of descent of $\varphi(z)$ follows from the equality $\psi'_k(z_k) = \varphi'(z_k)$, where

$$\begin{aligned} \varphi'(z_k) &= f_{0z}(x_k) + y'^*(z_k) f_{0y}(x_k), \\ f_{0z} &= \left(\frac{\partial f_0}{\partial x^{m+1}}, \dots, \frac{\partial f_0}{\partial x^n} \right), \\ f_{0y} &= \left(\frac{\partial f_0}{\partial x^1}, \dots, \frac{\partial f_0}{\partial x^m} \right), \\ y'(z_k) &= -g_y^{-1}(x_k) g_z(x_k). \end{aligned} \quad (9.10)$$

Since the process of the (9.9) type can be considered to be a method of minimization of function $\varphi(z)$, it is easy to see that vector $\varphi'(z_k)$ can be taken as p_k ; then sequence (9.9) will be the gradient method of minimization of $\varphi(z)$.

Note that vector p_k providing the minimum of function $F_k(z, y_l(z))$, where $F_k(z, y)$ is defined by expression (9.5), is calculated by the following formula:

$$p_k = -(I + y'^*(z_k) y'(z_k))^{-1} \varphi'(z_k). \quad (9.11)$$

Consequently, sequence (9.9) in which p_k is determined by formula (9.11) is also a method of the gradient type for the minimization of $\varphi(z)$. *Methods of the gradient type* we shall call *methods of the first order*.

Newton's method and its modifications can be applied in principle to the minimization of function $\varphi(z)$, provided certain necessary requirements are fulfilled. However, it should be noted that the calculation of the second derivative $\varphi''(z)$ is, as a rule, very laborious, for it requires the calculation of the second derivative of the vector-function $y(z)$, i.e., practically, the calculation of the second derivatives of functions $f_1(x), \dots, f_m(x)$.

Suppose that condition (9.8) is not fulfilled and the following weaker requirement is satisfied: at any point $x \in S_g$ at least one determinant of order m is not zero

$$\left| \left\{ \frac{\partial f_i}{\partial x^j} \right\} \right| \neq 0, \quad (9.12)$$

$$j = j_1, j_2, \dots, j_m, \quad j \in [1, \dots, n], \quad i = 1, \dots, m.$$

The weakening of the requirements to functions f_i is that at different points of set S_g different determinants may be not zero. In this case, the coordinates of point $x \in S_g$ which form vector z and vector-function $y(z)$ can be, speaking generally, different at different points of manifold $S_g : z = (x^{j_{m+1}}, \dots, x^{j_n}), y = (x^{j_1}, \dots, x^{j_m})$. Taking this into account we can, as before, use formula (9.9) for the restoration of ties. Each step of process (9.9) can be treated as a step of the process of minimization of a certain function $\varphi(x^{j_{m+1}}, \dots, x^{j_n})$ for which the corresponding vector p_k is the direction of descent.

Methods of the (9.9) type will be studied below. It will be convenient to denote any vector-function $(x^{j_1}, \dots, x^{j_m})$ by y and a vector of independent variables by z (as we did in fulfilling the condition (9.8)). Accordingly, any of the determinants $|\{\partial f_i / \partial x^j\}|$ of m order will be denoted by $|g_y|$ and function $f_0(z, y(z))$ by $\varphi(z)$. The absolute value of function $|g_y(x)|$ shall be denoted by $|g_y(x)|_A$.

In the following two subsections we shall study the properties of the methods of the first and second order. In the fourth subsection we shall consider methods of dual and conjugate directions for the minimization of $\varphi(z)$ (or the algorithms based on methods of this type). From the viewpoint of practical computations just these algorithms are of the greatest interest.

Methods of the First Order

We shall study the properties of methods based on the linearization of function $f_0(x)$ and ties f_i , $i = 1, \dots, m$.

Consider the algorithm whose every step is a step of the gradient method for minimization of a certain function $\varphi(z)$:

$$z_{k+1} = z_k - \alpha_k \varphi'(z_k), \quad y_{k+1} = y(z_{k+1}) \quad (9.13)$$

where z_k is a vector corresponding to the determinant $|\bar{g}_y(x)|$ which has at point $x_k \in S_g$ the maximum absolute value of all the determinants $|g_y|$, the gradient $\varphi'(z_k)$ is calculated by formula (9.10) and parameter α_k can be determined by one of the methods described in studying gradient methods (Sec. 1, Chap. II). We shall choose as α_k the maximum value of the parameter obtained by successive reductions of a certain positive constant which satisfies

$$f_0(z, y(z)) - f_0(z_k, y_k) \leq -\varepsilon \alpha \|\varphi'(z_k)\|^2, \quad 0 < \varepsilon < 1 \quad (9.14)$$

where $z = z_k - \alpha \varphi'(z_k)$ (this is an analogue of the method of choosing α_k according to condition (1.2), Chap. II).

Theorem 9.1. *If functions $f_0(x)$ and $f_i(x)$, $i = 1, \dots, m$ are twice continuously differentiable and, besides, functions f_i are such that condition (9.12) is fulfilled, and set $S = S_g \cap S_0$ ($S_0 = \{x: f_0(x) \leq f_0(x_0)\}$) is bounded with an arbitrary choice of point x_0 , then on sequence (9.13) $f_0(x_{k+1}) \leq f_0(x_k)$ and $\|\varphi'(z_k)\| \rightarrow 0$ as $k \rightarrow \infty$.*

Proof. The possibility of constructing sequence (9.13) follows from condition (9.12): with sufficiently small values of parameter α , point z_{k+1} lies in the region about point z_k , where function $y(z)$ is defined. In this region, by the assumptions of the theorem, function $\varphi(z) = f(z, y(z))$ is twice continuously differentiable. Taking this into account, the following estimate holds:

$$\varphi(z_{k+1}) - \varphi(z_k) \leq \alpha_k \|\varphi'(z_k)\|^2 \left(-1 + \frac{\alpha_k}{2} \|\varphi''(z_c)\| \right) \quad (9.15)$$

where $\varphi''(z_c) = \varphi''(z_k + \theta(z_{k+1} - z_k))$, $\theta \in [0, 1]$; hence, if the value of α_k is sufficiently small, the inequality (9.14) will be satisfied. This means that on elements of sequence (9.13) function $f_0(x)$ decreases monotonically.

We shall prove now that $\|\varphi'(z_k)\| \rightarrow 0$. On the closed bounded set S continuous function $|\bar{g}_y(x)|_A$ assumes its minimum value γ (Weierstrass' theorem), and by (9.12) $\gamma > 0$ (function $|\bar{g}_y(x)|_A$ is continuous being the maximum of continuous functions $|g_y(x)|_A$). The number of different functional determinants $|g_y|$ is finite and, since functions f_i are differentiable on set S , they are all uniformly continuous. Therefore, for any constant $0 < \gamma_1 < \gamma$, there is a con-

stant $\rho > 0$ such that at any point of set S which belongs to sphere S_θ of radius ρ and having its centre at an arbitrary point $\theta \in S$, the absolute value of determinant $|g_y(x)|$ assuming at point θ the value of $|\bar{g}_y(\theta)|$ will be not less than γ_1 . Moreover, since set S is bounded and the first and second partial derivatives of functions f_i are continuous, these derivatives at any point $\theta \in S$ chosen in sphere S_θ of radius ρ are bounded (by a constant M). Taking this statement into account and according to the theorems on implicit functions, we can assert that in a certain parallelepiped

$$[\theta^i \pm \delta\theta^i], \quad i = 1, \dots, n$$

belonging to sphere S_θ , system (9.1) defines at least one twice continuously differentiable vector-function $y(z)$ and in this parallelepiped the derivatives of any such function are bounded:

$$\|y'(z)\| \leq N_1, \quad \|y''(z)\| \leq N_1. \quad (9.16)$$

Since the derivatives $y'(z)$, $y''(z)$ are bounded and the first and second derivatives of function $f_0(x)$ on set S are bounded too, the derivatives of function $\varphi(z)$ in parallelepiped $[\theta^i \pm \delta\theta^i]$ are also bounded: $\|\varphi'(z)\| \leq N_2$, $\|\varphi''(z)\| \leq N_2$.

Taking into account the above remarks, we can ascertain that there is a constant $\delta > 0$ such that if $\alpha_k \leq \delta$ then point x_{k+1} is in sphere S_{x_k} of radius ρ . Indeed, suppose that $\|x_{k+1} - x_k\|^2 = \|z_{k+1} - z_k\|^2 + \|y_{k+1} - y_k\|^2 = \rho^2$. Since these equalities mean that $x_{k+1} \in S_{x_k}$, this point also belongs to parallelepiped $[x_k^i \pm \delta x_k^i]$, $i = 1, \dots, n$ in which estimates (9.16) hold for derivatives of function $y(z)$. Consequently, $\|y_{k+1} - y_k\| \leq N_3 \|z_{k+1} - z_k\|$ and therefore it follows from the preceding equalities that $(N_3^2 + 1) \|z_{k+1} - z_k\|^2 = \alpha_k^2 N_4^2 \|\varphi'(z_k)\|^2 \geq \rho^2$; hence

$$\alpha_k \geq \frac{\rho}{N_4 \|\varphi'(z_k)\|} \geq \frac{\rho}{N_4 N_2}.$$

This estimate shows that the equality $\|x_{k+1} - x_k\| = \rho$ holds with $\alpha_k \geq \rho/(N_4 N_2)$, i.e. that we can choose as δ any constant not exceeding $\rho/(N_4 N_2)$.

Using now inequality (9.15) (and taking into account that derivative $\varphi''(z)$ is bounded), it is easy to ascertain that inequality (9.14) will certainly hold with $\alpha_k = \min \left\{ \delta, \frac{2(1-\varepsilon)}{N_2} \right\}$. But this means, since $f_0(x)$ has a lower bound (on set S), that as $k \rightarrow \infty$ of necessity $\|\varphi'(z_k)\| \rightarrow 0$. The theorem is proved.

The condition $\|\varphi'(z_k)\| \rightarrow 0$ means in the general case that sequence (9.13) (or a certain of its subsequences) converges to point x_* which satisfies the necessary condition for an extremum of function $f_0(x)$ on manifold S_g (at point x_* the gradient $f'_0(x_*)$ is orthogonal

to the tangent hyperplane $g'(x_*)(x - x_*) = 0$ (Chap. I, Sec. 4). Since function $f_0(x)$ is continuous, its minimum on set S exists. If sequence (9.13) converges to the solution and function $\varphi(z)$, to whose minimization in a certain region about the minimum the solving of the original problem is reduced, satisfies the conditions $m_0 \|v\|^2 \leq (\varphi''(z)v, v) \leq M \|v\|^2$ for any $v \in E^{n-m}$, then the rate of convergence will not be slower than that of a certain geometric progression; this follows from the results on the convergence of gradient methods (theorem 1.2, Chap. II). We shall dwell on several questions connected with the implementation of the method.

In the theorem proved above, use was made in defining vectors z_k and y_k of the determinant $|\bar{g}_y(x_k)|$. To find this determinant it is necessary at each iteration to calculate all the determinants $|g_y|$. However in practice, there is no need to do so (the determinant $|\bar{g}_y|$ is used in the theorem only to simplify the proof). The convergence of the method is retained if we choose vectors z_k and y_k corresponding to any of the determinants $|g_y|$ whose absolute value at point x_k is not less than an arbitrary small positive constant μ (the same for all k). With the conditions of the theorem such a constant exists since there is the constant γ . Therefore in implementing the algorithm, vectors z and y should be chosen corresponding to the same determinant until at a certain point its absolute value becomes less than μ ; only if this has occurred is it necessary to pass to other vectors z and y , i.e. to calculate another determinant $|g_y|$. The constant μ is arbitrary. It can occur that at a certain point x_k all the determinants $|g_y|$ have an absolute value less than μ . We have then to choose a new constant $\mu_1 < \mu$. At each reduction of parameter α which is necessary to fulfill the inequality (9.14), a new evaluation of function $y(z)$ is required (for the evaluation of $f_0(x) = f_0(z, y(z))$), i.e. we have to solve the system of nonlinear equations (9.1) with a fixed value of vector z . To reduce the amount of computations, the required value of the parameter should be determined by establishing the following inequality:

$$f_0(z, y_l(z)) - f_0(z_k, y_k) \leq -\varepsilon \alpha \|\varphi'(z_k)\|^2, \quad 0 < \varepsilon < 1. \quad (9.17)$$

As soon as this inequality is satisfied, the inequality (9.14) should be checked with the α obtained; if (9.14) is not satisfied, the reduction of α should be continued; otherwise the obtained value of the parameter should be retained or one should attempt to increase it in checking (9.14). Note that

$$\begin{aligned} y(z) &= y_k + y'(z_k)(z - z_k) + O(\|z - z_k\|^2) \\ &= y_l(z) + O(\|z - z_k\|^2). \end{aligned}$$

With a sufficiently small $z_{k+1} - z_k$, we obtain $f_0(z_{k+1}, y_l(z_{k+1})) \rightarrow f_0(z_{k+1}, y(z_{k+1}))$; therefore if (9.17) is satisfied, inequality (9.14) will also be satisfied, i.e. no additional reductions of the step length will be required.

Remark. The requirements of theorem 9.1 to the smoothness of functions $f_0(x)$ and $f_i(x)$ can be taken somewhat weaker; however, this leads to a more complicated proof.

We dwell briefly also on a method of the (9.9) type in which vector p_k is chosen by formula (9.11) (vectors z_k, y_k are determined in the same way as in the preceding method) and α_k is the maximum value of the parameter (obtained by reductions) which satisfies the following inequality:

$$f_0(z, y(z)) - f_0(z_k, y_k) \leq \varepsilon \alpha (\varphi'(z_k), p_k), \quad z = z_k + \alpha p_k.$$

For such an algorithm, theorem (9.1) holds true. The proof will differ only in some details (analogously to the difference between the proof of the theorem of the properties of methods of the gradient type and the proof of the theorems on the method of steepest descent, Sec. 1, Chap. II).

Note that the amount of work per iteration in such an algorithm is greater than in method (9.13).

Method of the Second Order

Suppose that $f_0(x)$ is a strongly convex function. Then the quadratic function $F_k(x)$ (9.7) is strictly convex and since function $y_l(z)$ is linear, function $\psi_k(z) = F_k(z, y_l(z))$ is strictly convex too. More precisely, due to the strong convexity of $f_0(x)$, the following conditions are fulfilled for any function $\psi_k(z)$ with any vector $v \in E^{n-m}$:

$$m_0 \|v\|^2 \leq (\psi_k'' v, v) \leq M_0 \|v\|^2, \quad m_0 > 0 \quad (9.18)$$

where matrix $\psi_k'' = f_{0zz} + y' * f_{0yy} y' + 2y' * f_{0zy}$ (all the derivatives are calculated at point x_k). In this case vector p_k which minimizes $\psi_k(z)$ is calculated by the formula

$$p_k = -(\psi_k'')^{-1} \psi_k' (z_k). \quad (9.19)$$

In the method of the second order, point x_{k+1} , $k = 0, 1, \dots$, is constructed as follows:

$$z_{k+1} = z_k - \alpha_k (\psi_k'')^{-1} \psi_k' (z_k), \quad y_{k+1} = y(z_{k+1}) \quad (9.20)$$

where vectors z_k and y_k are determined in the same way, as in constructing method (9.13), and for α_k we take the maximum value of the parameter (obtained by reductions) which satisfies the follow-

ing inequality:

$$f_0(x) - f_0(x_k) \leq \varepsilon \alpha(\varphi'(z_k), p_k), \quad 0 < \varepsilon < \frac{1}{2} \quad (9.21)$$

where $x = (z, y(z))$, $z = z_k + \alpha p_k$.

Theorem 9.2. *Let $f_0(x)$ be a twice continuously differentiable function and for any vector $\omega \in E^n$*

$$m \|\omega\|^2 \leq (f_0''(x)\omega, \omega) \leq M_1 \|\omega\|^2, \quad m > 0$$

and let functions $f_i(x)$, $i = 1, \dots, m$ satisfy the requirements of theorem 9.1. Then, whatever the point x_0 chosen, the results of theorem 9.1 hold for method (9.20).

The proof of the theorem follows the same scheme as that used in proving theorem 9.1. Therefore, we shall dwell only on those changes in the proof, as compared to that of theorem 9.1, which arise because of different methods of choosing vector p_k .

Due to the strict convexity of $f_0(x)$, the set S_0 has a bound. It follows that set $S = S_g \cap S_0$ is bounded and closed (since sets S_0 and S_g are closed). Taking this into account, we prove, in the same way as in theorem 9.1, that estimates (9.16) hold and establish that the derivatives $\varphi'(z)$, $\varphi''(z)$ have bounds in the parallelepiped $[\theta^i \pm \delta\theta^i]$, $i = 1, \dots, n$.

Further, by (9.18), $\|(\psi_k'')^{-1}\| \leq \frac{1}{m_0}$, consequently, $\|p_k\| = \|(\psi_k'')^{-1} \psi_k'(z_k)\| = \|(\psi_k'')^{-1} \varphi'(z_k)\| \leq N_2/m_0$, and, therefore, if $(N_3^2 + 1) \|z_{k+1} - z_k\|^2 = \alpha_k^2 N_4^2 \|p_k\|^2 \geq \rho^2$, then

$$\alpha_k \geq \frac{\rho}{N_4 \|p_k\|} \geq \frac{\rho m_0}{N_4 N_2}. \quad (9.22)$$

It follows that any constant, provided it does not exceed $\rho m_0/(N_4 N_2)$, can be chosen as δ .

Using the expansion of function $\varphi(z)$ into Taylor's series, we find that in sphere S_{x_k} of radius ρ

$$\begin{aligned} \varphi(z_{k+1}) - \varphi(z_k) &= \alpha_k (\varphi'(z_k), p_k) + \frac{\alpha_k^2}{2} (\varphi''(z_k) p_k, p_k) \\ &\leq \alpha_k (\varphi'(z_k), p_k) \left(1 + \frac{\alpha_k}{2} \frac{N_2 \|p_k\|^2}{(\varphi'(z_k), p_k)} \right). \end{aligned}$$

It follows from (9.19), with account of (9.18), that

$$(\psi_k'' p_k, p_k) = -(\varphi'(z_k), p_k) \geq m_0 \|p_k\|^2.$$

This implies that

$$\varphi(z_{k+1}) - \varphi(z_k) \leq \alpha_k (\varphi'(z_k), p_k) \left(1 - \frac{\alpha_k}{2} \frac{N_2}{m_0} \right).$$

Taking into account this estimate and inequalities (9.22), we establish that inequality (9.21) will certainly hold with

$$\alpha_k = \min \left\{ \delta, \frac{2m_0(1-\varepsilon)}{N_2} \right\}.$$

This means, since $f_0(x)$ has a lower bound, that

$$(\varphi'(z_k), p_k) \rightarrow 0. \quad (9.23)$$

Since $-(\varphi'(z_k), p_k) = ((\psi_k'')^{-1}\varphi'(z_k), \varphi'(z_k)) \geq m_0 \|\varphi'(z_k)\|^2$, it follows from (9.23) that $\|\varphi'(z_k)\| \rightarrow 0$. This completes the proof of the theorem.

In implementing the algorithm (9.20) one should take account of the remarks concerning the choice of vectors z_k, y_k and parameter α_k made in studying method (9.13).

If sequence (9.20) converges to the solution and the condition

$$\psi_k'' \rightarrow \varphi''(z_k) \quad (9.24)$$

holds for function $\varphi(z)$ to whose minimization the solving of the original problem is reduced, then the rate of convergence of the method is superlinear. In order to ascertain this, one should take into account that if (9.24) holds (and also $(\psi_k'' p_k, p_k) = -(\varphi'(z_k), p_k)$), then

$$\begin{aligned} \varphi(z_{k+1}) - \varphi(z_k) &= \alpha_k (\varphi'(z_k), p_k) \\ &\times \left(1 - \frac{\alpha_k}{2} \frac{(\varphi''(z_k) p_k, p_k)}{(\psi_k'' p_k, p_k)} - \frac{\alpha_k}{2} \frac{((\varphi''(z_{kc}) - \varphi''(z_k)) p_k, p_k)}{(\psi_k'' p_k, p_k)} \right) \\ &\rightarrow \alpha_k (\varphi'(z_k), p_k) \left(1 - \frac{\alpha_k}{2} - \frac{\alpha_k}{2} \frac{((\varphi''(z_{kc}) - \varphi''(z_k)) p_k, p_k)}{(\psi_k'' p_k, p_k)} \right). \end{aligned}$$

At the same time, by (9.18) and (9.24), function $\varphi(z)$ will be strongly convex in a certain region about the minimum. With the above remarks, the proof of the superlinear rate of convergence can be the same as that, for example, in studying Newton's method (Sec. 2, Chap. II).

Thus the rate of convergence of method (9.20) in a number of problems will be faster than that of methods of the first order. However, the amount of work per iteration in method (9.20) may prove considerably greater owing to the necessary calculations of the second derivatives of function $f_0(x)$.

Minimization Methods of Higher Effectiveness

The projection methods described in the preceding subsections are, in a sense, analogues of the gradient methods and Newton's method for solving problems of finding an absolute extremum.

They share the shortcomings of the corresponding methods: either they have a slow rate of convergence (like methods of the first order) or they involve a greater amount of work per iteration (like methods of the second order). However, the fact that in the algorithm under consideration the solving of the original problem is reduced to unconstrained minimization of a function (one or several functions, depending on whether condition (9.8) or (9.12) holds) makes it possible to use such effective minimization algorithms as methods of dual and conjugate directions (Secs. 3-5, Chap. II). Thus, if functions $f_i(x)$, $i = 1, \dots, m$ are such that condition (9.8) is satisfied and with any fixed z system (9.1) has a unique solution $y = y(z)$, then, if $\varphi(z) = f(z, y(z))$ is a twice continuously differentiable strongly convex function, any method of dual or conjugate directions (for the minimization of $\varphi(z)$) converges to the solution at a superlinear rate. At the same time if use is made of the variants of methods with restoration of matrices A_k^{-1} and H_k after a finite number of steps (see the subsection on p. 104), then the convergence of methods of dual and conjugate directions will be guaranteed with the same assumptions about function $\varphi(z)$ as in gradient methods.

Consider, for example, the problem of the minimization of quadratic function $f_0(x)$ with linear constraints: $g(x) \equiv Ax + b = 0$, where $A = (a_{ij})$ is an $m \times n$ matrix, $b = (b^1, \dots, b^m)$.

Let the determinant $|a_{il}| \neq 0$, $i, l = 1, \dots, m$; then we can take $y = (x^1, \dots, x^m)$, $z = (x^{m+1}, \dots, x^n)$.

Since $y(z)$ is a linear function, $\varphi(z)$ is a quadratic function of the variable z and it is strictly convex if the original function $f_0(x)$ is strictly convex. The application of any method of dual or conjugate directions makes it possible to find the minimum of function $\varphi(z)$ after $n - m$ steps.

If function $f_0(x)$ and $f_i(x)$, $i = 1, \dots, m$ satisfy the requirements of theorem 9.2 (the condition (9.8) in this case is substituted by the weaker requirement (9.12)), then methods of dual or conjugate directions can be used to minimize each of the functions $\varphi(z)$ that are to be dealt with in solving the problem. In other words, in algorithms of the (9.9) type, vector p_k and parameter α_k can be determined in the same way as in dual or conjugate directions methods, and vectors z_k and y_k can be chosen as in studying method (9.13). Algorithms thus constructed (implemented with restoration of matrix A_k^{-1} or H_k) converge under the same conditions as projection methods of the first and second orders. At the same time, their effectiveness is higher than that of the projection methods studied; in particular, with a small increase in the amount of work per iteration as compared to that in the method of the first order, a superlinear rate of convergence can be attained.

In practice, just the algorithms of the type described in this subsection should be used for solving the problem being studied.

Note also that the amount of work per iteration (method (9.20)) can be reduced if instead of matrix $f_0''(x_k)$ use is made of matrix D_k defined by the following system of equations:

$$D_k(x_{k-i} - x_{k-i-1}) = f_0'(x_{k-1}) - f_0'(x_{k-i-1}),$$

$$i = 0, \quad 1, \dots, \quad n-1$$

(analogue of system (3.6) of Chap. II which is used in constructing methods of dual direction) and vector $p_k = -F_k^{-1}\varphi'(z_k)$ is constructed, where $F_k = D_{kzz} + y'^*D_{kyy}y' + 2y'^* + D_{kzy}$ and matrices D_{kzz} , D_{kyy} , D_{kzy} are parts of matrix D_k and correspond to matrices f_{0zz} , f_{0yy} , f_{0zy} , respectively.

On the Solving of the General Problem of Mathematical Programming

It is required to minimize function $f_0(x)$ with constraints

$$f_i(x) \leq 0, \quad i = 1, \dots, m. \quad (9.25)$$

Such constraints can be reduced to equality constraints in several ways. For instance, if we introduce additional variables $x^{n+1}, \dots, x^{n+m}$, then constraints (9.25) will hold with the same values of variables $x^1, \dots, x^n$ which satisfy the equalities

$$(x^{n+i})^2 + f_i(x) = 0, \quad i = 1, \dots, m. \quad (9.26)$$

Consequently, the minimum of function $f_0(x)$ with constraints (9.25) will coincide with the minimum of $f_0(x)$ with constraints (9.26). For the minimization of $f_0(x)$ with constraints (9.26), methods of the first order described in the subsection on p. 249 can be used.

Method (9.20) with conditions (9.26) cannot be applied to the minimization of $f_0(x)$, for in space E^{n+m} function $f_0(x)$ is not strictly convex; matrix $f_0''(x)$, as is easily ascertained, is singular in E^{n+m} . Due to this fact, there is no point in using methods of dual and conjugate directions in this case.

Conclusive Remarks

Of the class of projection methods with restoration of ties we have discussed only those algorithms in which use is made of formulas (9.9) to restore ties. In a number of problems this method of carrying out the concluding (third) stage of iteration may prove inconvenient; in this case it is worth while to carry out this stage in another way. For instance, one can determine point $x_{k+1} \in S_g$ by making the quantity $\|x_{k+1}(\alpha) - \hat{x}_k(\alpha)\|$ minimize the distance between point $\hat{x}_k(\alpha)$ and set S_g .

The amount of work per iteration in projection methods diminishes when approaching the solution (or a stationary point of function $f_0(x)$ on S_g) since the problem of determining point x_{k+1} (using point $\hat{x}_k$) becomes simpler. So, for example, the amount of work involved in solving system (9.1) with a fixed vector z_{k+1} (i.e. in evaluating function $y(z_{k+1})$) diminishes as we approach the solution of the problem because point $y_l(z_{k+1})$ with increasing k approximates the solution $y(z_{k+1})$ better and better. In this sense, projection methods differ advantageously from penalty function methods where we have to solve problems of greater and greater complexity in order to obtain a more precise approximation to the solution.

Bibliographic Notes

To Sec. 1. The method of solving problems of quadratic programming expounded in this section is based on applying the method of conjugate gradients. It is the most simple and expedient method if the constraints on the variables are simple. There are a great many other methods of solving the problem of quadratic programming which converge either after a finite number or an infinite number of steps. These methods are described by H. Künzi and W. Krelle, S. I. Zukhovitskii and L. I. Avdeyeva, G. Zoutendijk [1], V. F. Dem'yanov and V. N. Malozemov [2].

The problems of the effectiveness and accuracy of different algorithms are analyzed by V. V. Ivanov [2], [3], V. V. Ivanov and V. E. Truten'.

To Sec. 2. In describing the method of feasible directions we followed mainly the works by S. I. Zukhovitskii, R. A. Polyak and M. E. Primak [1], [2]. The method described in Sec. 2 differs from the traditional one by the rule of choosing the step length at each iteration. Several variants of the method of feasible directions have been studied in detail and substantiated by G. Zoutendijk [1], and also in papers by G. Zoutendijk [2], D. M. Topkis and A. Veinott.

To Sec. 3. The method of conditional gradient was for the first time described by M. Frank and P. Wolfe. Later on it was studied by V. F. Dem'yanov and A. M. Rubinov, E. S. Levitin and B. T. Polyak who gave the bounds on the rate of convergence. The proof of the accuracy of these estimates is given in a paper by M. D. Canon and C. D. Cullum.

The generalization of Newton's method of solving constrained problems was carried out by E. S. Levitin and B. T. Polyak. Newton's method with step adjustment with constraints was studied by Yu. M. Danilin [1], [2].

To Sec. 4. The cutting hyperplane method as described in this section follows the work by J.E. Kelley. Different generalizations of this method and also the bounds on its rate of convergence ($|f(x_k) - f(x_*)| \leq C/n$) for a strongly convex function $f(x)$ are given in the paper by E. S. Levitin and B. T. Polyak.

To Secs. 5, 6. The description of the linearization method in this chapter follows the work by B. N. Pshenichny [4] where he proves that the method is convergent. The subtler results such as the finite convergence in linear programming, the local estimate of the rate of convergence and also quadratic rates of convergence in special cases, had not been published before. The same can be said of the application of this method to the problem of finding the minimax. The last problem was studied by V. F. Dem'yanov and V. N. Malozemov [1], [2] who constructed a number of algorithms of descent for solving the problem.

Note also that the minimax problem can be solved by using the algorithm of the generalized gradient descent and its variants developed by N. Z. Shor [1]-[4].

To Sec. 7. The methods of the acceleration of convergence described on pp. 225-233 had not been published before; however, in its basic ideas it is closely connected with the works by Yu. M. Danilin and B. N. Pshenichny [1], [2]. Methods of solving systems of equations without using derivatives of the left-hand sides were described in many works; of these we mention here only the works by F. J. Zeleznik, J. Barnes, C. G. Broyden [1], V.E. Shamansky [1].

The application of Newton's method of solving systems of equations to those arising in formulating the necessary conditions for an extremum is described in the subsection on p. 233. The problems involved are treated in detail by B. T. Polyak [3].

To Sec. 8. The literature on the very popular penalty function method is extensive. The different properties of this method can be found in monographs by J. Céa, A. V. Fiacco and G. P. McCormick, E. Polak [2], in papers by W. I. Zangwill [2], E. S. Levitin and B. T. Polyak, A. V. Fiacco. The bounds on the rate of convergence were studied by I. I. Eremin [1], D. Laenberger. The method of centers closely connected with the penalty function method was suggested by P. Huard. A new, little studied variant of the penalty function method is given in a work by M. R. Hestenes [1].

To Sec. 9. One of the first authors to begin using projection methods for solving constrained problems was J. B. Rosen. In a paper by A. Miele, H. Y. Huang and J. C. Heideman projection methods with restoration of ties are described in which use is made of the gradient method or the conjugate gradient method for determining the direction of movement and at the conclusive stage the problem of the minimization of the distance between the point and the tie is solved.

The projection algorithms considered in this section were described by Yu. M. Danilin [3]. Finally, we mention here several of the methods that have not been described in this book: the method of Fejer's approximations developed by I. I. Eremin [2], [3], and also N. Z. Shor's method of the generalized gradient [1]-[4]. Combined methods of finding extrema were studied by V. V. Ivanov [1]. Many interesting results on the convergence of minimization algorithms have been obtained by V. G. Karmanov.

Of papers devoted to reviews of numerical methods we suggest those by Yu. M. Ermol'ev, G. Zoutendijk [2], and also books by H. Künzi and W. Oettli, F. P. Vasil'ev; the latter contains an extensive bibliography on the problem under consideration.

APPENDIX

COMPUTATIONAL SCHEMES OF THE MAIN ALGORITHMS

I. METHOD OF DUAL DIRECTIONS (CHAP. II, SEC. 3)

This method is intended for the minimization of a convex function $f(x)$, $x \in E^n$.

Iteration scheme.

Let x_0 be an arbitrary point, $s_{0,0}, s_{0,-1}, \dots, s_{0,-n+1}$ be an arbitrary linearly independent vector system.

With $0 \leq k \leq n-1$, the iteration is as follows:

(1) Construct the point

$$x_{k+1} = x_k - \alpha_k f'(x_k) \quad (1)$$

where α_k is chosen by any of the methods described in Chap. II, Sec. 1.

(2) Set:

$$\begin{aligned} r_{k+1} &= x_{k+1} - x_k, \\ e_{k+1} &= f'(x_{k+1}) - f'(x_k). \end{aligned} \quad (2)$$

(3) Compute

$$(s_{k,k-n+1}, e_{k+1}).$$

If

$$|(s_{k,k-n+1}, e_{k+1})| \geq \gamma \|s_{k,k-n+1}\| \|e_{k+1}\| \quad (3)$$

where $\gamma > 0$ is an arbitrarily small constant, go to step (5).

If

$$|(s_{k,k-n+1}, e_{k+1})| < \gamma \|s_{k,k-n+1}\| \|e_{k+1}\|, \quad (4)$$

go to step (4).

(4) Set

$$r_{k+1} = \beta_{k+1} s_{k,k-n+1} \quad (5)$$

where the quantity $\beta_{k+1} > 0$ is chosen such that the condition $\|r_{k+1}\| < \|r_k\|$ be fulfilled.

Compute the gradient $f'(x_k + r_{k+1})$ and then construct vector $e_{k+1} = f'(x_k + r_{k+1}) - f'(x_k)$. Then go to step (5).

(5) Construct the vector system

$$\begin{aligned} s_{k+1,k+1} &= \frac{s_{k,k-n+1}}{(s_{k,k-n+1}, e_{k+1})}, \\ s_{k+1,k-j} &= s_{k,k-j} - (s_{k,k-j}, e_{k+1}) s_{k+1,k+1}, \\ j &= 0, 1, \dots, n-2. \end{aligned} \quad (6)$$

This is the end of the iteration.

With $k \geq n$:

(1) Construct vector

$$p_k = - \sum_{i=0}^{n-1} (f'(x_k), s_{k, k-i}) r_{k-i}.$$

(2) Compute $(f'(x_k), p_k)$.

If $(f'(x_k), p_k) \leq 0$, construct point x_{k+1} by one of the formulas $x_{k+1} = x_k \pm \alpha_k p_k$, where α_k is chosen according to condition (2.2), Chap. II.

If $(f'(x_k), p_k) = 0$, construct point x_{k+1} using the gradient method (see (1)) of the iteration with $k \leq n-1$ (further, construct the iteration in the same way as you did with $k \leq n-1$ (steps (2)-(5)).

Remarks. 1. We have adduced only one of the possible computation schemes of the methods of dual directions. Here the first iterations ($k \leq n-1$) are carried out by the gradient method. Since at the initial steps of the iterative process the gradient method usually provides for a sufficiently steep decrease of the function, such an initial stage of the process is expedient in solving many problems.

2. We have changed here for the sake of convenience the notations of the vectors of the dual basis (cf. (6), and (3.21) of Chap. II).

3. If vector r_{k+1} is chosen in form (5) and function $f(x)$ is smooth and strongly convex (i.e. conditions (2.4) of Chap. II hold true), then inequality (3) will automatically hold, provided the constant γ has been chosen sufficiently small. Indeed, if function $f(x)$ satisfies the requirements formulated, then $\|e_k\| \leq M \|r_k\|$ and estimate (5.18) of Chap. II holds true. By virtue of this fact, we have

$$\begin{aligned} (s_{k, k-n+1}, e_{k+1}) &= \frac{1}{\beta_k} (r_{k+1}, e_{k+1}) \geq \frac{m}{\beta_k} \|r_{k+1}\|^2 \\ &\geq \frac{m}{\beta_k M} \|r_{k+1}\| \|e_{k+1}\| = \frac{m}{M} \|s_{k, k-n+1}\| \|e_{k+1}\|. \end{aligned}$$

Thus if $\gamma \leq \frac{m}{M}$, then inequality (3) is satisfied.

4. The practice of computations shows that quantity γ may be chosen very small: $\gamma = 10^{-6}$ - 10^{-15} . If condition (3)₁ is not satisfied even with vector r_{k+1} having been chosen in form (5), it means that matrix $f''(x)$ as $x \rightarrow x_*$ becomes ill conditioned, i.e. the minimized function is not strongly convex. In particular, the surfaces of the levels of this function may have forms of a long, deep and narrow valley. In this case, a very accurate approximation to the solution with respect to the variable is not attainable. However, the computation practice shows that one can obtain function values sufficiently close to the minimum in minimizing even of nonconvex functions with deep valley level surfaces.

II. CONJUGATE GRADIENTS METHOD (CHAP. II, SEC. 4)

This method is intended for the minimization of convex function $f(x)$, $x \in E^n$. Iteration scheme.

Let x_0 be an arbitrary point, $p_0 = -f'(x_0)$.

(1) If $\begin{cases} 0 \leq k \leq n-1, & \text{go to (2).} \\ k = n, & \text{go to (5).} \end{cases}$

(2) Construct point

$$x_{k+1} = x_k + \alpha_k p_k$$

where factor α_k is determined under the condition

$$f(x_k + \alpha_k p_k) = \min_{\alpha \geq 0} f(x_k + \alpha p_k).$$

(3) Compute vector

$$p_{k+1} = -f'(x_{k+1}) + \beta_{k+1} p_{k+1}$$

where

$$\beta_{k+1} = - \frac{(f'(x_{k+1}), f'(x_{k+1}) - f'(x_k))}{(f'(x_k), p_k)}.$$

(4) Go to (1).

(5) Set $x_n = x_0$, $p_0 = -f'(x_n)$ and repeat the process (go to (1)).

Remark. Coefficient β_{k+1} can be determined by any of the formulas (4.73), Chap. II.

III. METHOD OF FEASIBLE DIRECTIONS (CHAP. III, SEC. 2)

This method is intended for solving problems of convex programming: to minimize function $f_0(x)$ with constraints

$$\begin{aligned} f_i(x) &\leq 0, \quad i = 1, \dots, m, \\ Ax - b &= 0 \end{aligned}$$

where $x \in E^n$, $f_i(x)$, $i = 0, \dots, m$ are convex continuously differentiable functions, A is an $l \times n$ matrix, b is an l -dimensional vector.

Notations:

$$\mathcal{J}_{\bar{\delta}}(x) = \{i : f_i(x) \geq -\delta, i = 1, \dots, m\},$$

$$\|p\| = \max_{1 \leq j \leq n} |p^j|$$

where $p \in E^n$, p^j are components of vector p .

Initial data: x_0 is the initial approximation satisfying all the constraints; $\delta_0 > 0$, $\xi_i > 0$, $i = 0, \dots, m$ are positive numbers which, speaking generally, are arbitrary.

The common step of the algorithm.

Point x_k and number $\delta_k > 0$ have been computed.

(1) Solve the problem of linear programming

$$\min \eta$$

$$(f'_i(x_k), p) \leq \xi_i \eta, \quad i \in \mathcal{J}_{\bar{\delta}_k}(x_k) \cup \{0\},$$

$$Ap = 0,$$

$$-1 \leq p^j \leq +1, \quad j = 1, \dots, n.$$

The solution is η_k, p_k .

(2) If $\eta_k < -\delta_k$, then

$$x_{k+1} = x_k + \alpha_k p_k, \quad \delta_{k+1} = \delta_k$$

where $\alpha_k = \frac{1}{2^{q_0}}$, and q_0 is the first integer of $q=0, 1, \dots$, for which the following inequalities hold:

$$f_0 \left(x_k + \frac{1}{2^q} p_k \right) \leq f_0(x_k) + \frac{1}{2} \frac{1}{2^q} \eta_k,$$

$$f_i \left(x_k + \frac{1}{2^q} p_k \right) \leq 0, \quad i=1, \dots, m.$$

(3) If $\eta_k \geq -\delta_k$, then $x_{k+1} = x_k$, $\delta_{k+1} = \frac{1}{2} \delta_k$.

(4) Return to (1).

Remark. The choice of numbers δ_0 , ξ_i can influence the course of the process; the choice should be made on the basis of an analysis of the problem under consideration. The algorithm can be used also for nonconvex problems.

IV. LINEARIZATION METHOD (CHAP. III, SEC. 5)

This method is intended for solving the problem: to minimize $f_0(x)$ with constraints:

$$\begin{aligned} f_i(x) &\leq 0, & i=1, \dots, m, \\ f_i(x) &= 0, & i=m+1, \dots, m+l \end{aligned}$$

where $f_i(x)$ are continuously differentiable functions.

Notations:

$$F(x) = \max \{0, f_1(x), \dots, f_m(x), |f_{m+1}(x)|, \dots, |f_{m+l}(x)|\},$$

$$\mathcal{J}_{\delta}^{-}(x) = \{i: f_i(x) \geq F(x) - \delta, i=1, \dots, m\},$$

$$\mathcal{J}_{\delta}^0(x) = \{i: |f_i(x)| \geq F(x) - \delta, i=m+1, \dots, m+l\},$$

$$\Phi_N(x) = f_0(x) + NF(x),$$

$$\|p\|^2 = \sum_{j=1}^n (p^j)^2.$$

Initial data: the initial approximation x_0 is arbitrary; N_0 is sufficiently great, $\delta_0 > 0$, $0 < \varepsilon < 1$.

The common step of the algorithm.

Point x_k is constructed and numbers N_k and δ_k are chosen.

(1) Solve the problem:

$$\min (f'_0(x_k), p) + \frac{1}{2} \|p\|^2,$$

$$(f'_i(x_k), p) + f_i(x_k) \leq 0, \quad i \in \mathcal{J}_{\delta_k}^{-}(x_k),$$

$$(f'_i(x_k), p) + f_i(x_k) = 0, \quad i \in \mathcal{J}_{\delta_k}^0(x_k).$$

The solution is p_k . If the problem is incompatible, then set $x_{k+1} = x_k$, $\delta_{k+1} = \frac{1}{2} \delta_k$, $N_{k+1} = N_k$ and return to (1).

(2) If the problem is consistent and p_k is found, then set

$$\begin{aligned}x_{k+1} &= x_k + \alpha_k p_k, \\ \delta_{k+1} &= \delta_k\end{aligned}$$

where α_k is chosen equal to $\frac{1}{2^{q_0}}$ and q_0 is the first of integers of $q = 0, 1, \dots$, for which the relation

$$\Phi_{N_k} \left(x_k + \frac{1}{2^q} p_k \right) \leq \Phi_{N_k}(x_k) - \frac{1}{2^q} \varepsilon \|p_k\|^2$$

holds.

(3) Let numbers u_k^i , $i \in \mathcal{J}_{\delta_k}^-(x_k) \cup \mathcal{J}_{\delta_k}^0(x_k)$ be Lagrange multipliers of the subsidiary problem that was solved at the first stage. If now

$$N_k > \sum_{i \in \mathcal{J}_{\delta_k}^-(x_k)} u_k^i + \sum_{i \in \mathcal{J}_{\delta_k}^0(x_k)} |u_k^i|,$$

then $N_{k+1} = N_k$.
Otherwise

$$N_{k+1} = 2 \left(\sum_{i \in \mathcal{J}_{\delta_k}^-(x_k)} u_k^i + \sum_{i \in \mathcal{J}_{\delta_k}^0(x_k)} |u_k^i| \right).$$

(4) Return to (1).

Remark. Numbers δ_k and N_k cease to change from a certain step on. The algorithm requires an effectively working standard program for solving the problem of quadratic programming.

V. ALGORITHM FOR SOLVING A SYSTEM OF EQUATIONS WITHOUT CALCULATING DERIVATIVES (CHAP. III, SEC. 6)

This algorithm is intended for solving the system of equations

$$p(x) = 0$$

where $x \in E^n$, $p(x)$ is an n -dimensional vector-function whose components $p^j(x)$, $j = 1, \dots, n$ are differentiable.

Initial data: initial approximations $x_1, \dots, x_n$ are arbitrarily chosen in a sufficiently small region about the solution. In a particular case all x_k , $k = 1, \dots, n$ can coincide.

Notations: $\|p(x)\|^2 = \sum_{j=1}^n (p^j(x))^2$; $\varphi(k)$ is equal to $1, 2, \dots, n-1$ if k

when divided by n leaves a remainder $1, 2, \dots, n-1$ respectively, $\varphi(k) = n$ if it divides by n .

The common step of the algorithm, $x_1, \dots, x_k$ have been constructed.

(1) Solve for unknowns β_i , $i = 1, \dots, n$ the system of equations

$$\sum_{i=1}^n s_{k-n+i} \beta_i = -p(x_k)$$

where

$$z_j = \frac{1}{\|p(x_j)\|} [p(x_j) + \|p(x_j)\| e_{\varphi(j)} - p(x_j)],$$

e_i is a vector with zero components, except the i -th one which is equal to 1.

(2) Set

$$x_{k+1} = x_k + \sum_{i=1}^n \beta_i e_{\varphi(k-n+i)}.$$

(3) Return to (1).

LITERATURE

- Altman, M. "Generalized gradient methods of minimizing a functional", *Bull. Acad. Polon. Sci.* 14, No. 6 (1966), 313-318.
- Arrow, K., Hurwiz, L., and Uzawa, H. (eds.) *Studies in Linear and Nonlinear Programming*, Stanford, Calif., 1958.
- Auslender, A. "Méthodes et théorèmes de dualité", *Comptes Rendus, Académie Science, Paris*, A267 (1968), 114-117.
- Balakrishnan, A. V., and Neustadt, L. W. (eds.) *Computing Methods in Optimization Problems*, New York, 1964.
- Barnes, J. "An algorithm for solving nonlinear equations based on the secant method", *Computer J.* 8 (1965) 66-72.
- Bershchansky, Ya. M. "A method for solving problems of linear and convex programming", *Zh. Vychisl. Mat. i Mat. Fiz.* 10, No. 3 (1970), 621-629 (in Russian).
- Bourbaki, N. *Eléments de mathématique. Livre 5. Espaces vectoriels topologiques*, Paris, 1953.
- Brent, R. P. "Algorithms for finding zeros and extrema of functions without calculating derivatives", *STAN-CS*, 71-198, Feb. 1971.
- Brown, K. M., and Dennis, J. E. "On Newton-like iteration functions: general convergence theorems and a specific algorithm", *Num. Math.* 12 (1968).
- Broyden, C. G. 1. "A class of methods for solving nonlinear simultaneous equations", *Math. Comp.* 19 (1965), 577-593.
 2. "Quasi-Newton methods and their application to function minimization", *Math. Comp.* 21 (1967), 368-381.
 3. "The convergence of single-rank quasi-Newton methods", *Math. Comp.* 24 (1970).
- Budak, B. M., and Goldman, N. R. "On the application of the Newton method for solving nonlinear boundary-value problems" in *Vychislitel'nye metody i programmirovaniye* (Computational Methods and Programming), Issue VI, Moscow University Press, 1967, pp. 17-33 (in Russian).
- Canon, M. D., and Cullum, C. D. "A tight upper bound on the rate of convergence of the Frank-Wolfe algorithm", *SIAM J. Control* 6, (1968), 509-516.
- Canon, M. D., and Eaton, J. H. "A new algorithm for a class of quadratic programming problems with application to control", *SIAM J. Control* 4, No. 1 (1966), 34-45.
- Cauchy, A. L. "Méthode générale pour la résolution des systèmes d'équations simultanées", *Comptes Rendus, Académie Science, Paris*, 25 (1847), 536-538.
- Céa, J. *Optimization théorie et algorithmes*, Paris, 1971.
- Collatz, L. *Functional Analysis and Numerical Mathematics*, New York, 1966.
- Daniel, J. W. 1. *The Approximate Minimization of Functionals*, Englewood Cliff, N. J., 1971.
 2. "Convergence of the conjugate gradient method with computationally efficient modifications", *Num. Math.* 10, No. 2 (1967), 125-131.
 3. "The conjugate gradient method for linear and nonlinear operator equations", *SIAM J. Num. Anal.* 4 (1967), 10-28.
- Danilin Yu. M. 1. "On an approach to minimization problems", *Doklady Akad. Nauk SSSR* 188, No. 6 (1969), 1221 [Engl. trans.: *Soviet Math. Doklady* 10 (1969), 1274].

2. "Minimization methods based on approximation of the initial functional by a convex one", *Zh. Vychisl. Mat. i Mat. Fiz.* 10, No. 5 (1970), 1067-1080 (in Russian).
 3. "On function minimization in problems with equality constraints" *Kibernetika*, No. 2 (1971), 88-95 (in Russian).
 4. "Methods of conjugate directions for solving minimization problems", *Kibernetika*, No. 5 (1971), 122-136 (in Russian).
- Danilin Yu. M. and Pshenichny B. N. 1. "On minimization methods with accelerated convergence", *Zh. Vychisl. Mat. i Mat. Fiz.* 10, No. 6 (1970), 1341-1354 (in Russian).
2. "Minimization method without calculating derivatives", *Zh. Vychisl. Mat. i Mat. Fiz.* 11, No. 1 (1971), 12-21.
- Dantzig G. B. 1. *Linear Programming and Extensions*, Princeton, 1963.
2. "Linear control processes and mathematical programming", *SIAM J. Contr.* 4, No. 1 (1966), 56-60.
- Davidon W. C. 1. "Variable metric methods for minimization", *AEC Research and Development, Rept. ANL 5990* (Rev.), 1959.
2. "Variance algorithm for minimization", *Comp. J.* 10, No. 4 (1968), 406-410.
- Dem'yanov V. F. and Malozemov V. N. 1. "Contribution to the theory of non-linear minimax problems", *Uspekhi Mat. Nauk* 26, No. 3 (1971), 53-104 (in Russian).
2. *Introduction to Minimax*, New York, 1974.†
- Dem'yanov V. F. and Rubinov A. M. *Approximate Methods of Optimization Problems*, New York, 1970.
- Dennis I. E. "On Newton-like methods", *Numer. Math.* Band 11, Heft 4 (1968).
- Dubovitsky A. Ya. and Milyutin A. A. "Extremal problems with constraints", *Zh. Vychisl. Mat.* 5, No. 3 (1965), 395-453 (in Russian).
- Dunford N. and Schwartz J. T. *Linear Operators. Part 1: General Theory*, New York, 1962.
- Eremin I. I. 1. "Penalty method in convex programming", *Kibernetika*, No. 4 (1967), 63-67 (in Russian).
2. "Method of Fejer approximations in convex programming", *Matem. Zametki* 3 (1968), 217-234 (in Russian).
 3. "Rate of convergence of the method of Fejer approximations", *ibid.* 4 (1968), 53-61 (in Russian).
- Ermol'ev Yu. M. "Methods for solving nonlinear extremal problems", *Kibernetika*, No. 4 (1966), 1-17 (in Russian).
- Faddeev D. K. and Faddeeva V. N. *Computational Methods of Linear Algebra*, San Francisco, 1963.
- Fiacco A. V. "Penalty method for mathematical programming in E^n with general constraint sets", *J. Opt. Theory and Appl.* 6, No. 3 (1970), 252-268.
- Fiacco A. V. and McCormick G. P. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, New York, 1968.
- Fikhtenhol'ts G. M. *Course of Differential and Integral Calculus*, Moscow, 1959 (in Russian).
- Fletcher R. 1. "A new approach to variable metric algorithms", *Comput. J.* 13, No. 3 (1970), 317-322.
2. "Function minimization without evaluating derivatives. A review", *Comput. J.* 8, No. 1 (1965), 33-41.
- Fletcher R. and Powell M. J. D. "A rapidly convergent descent method for minimization", *Comput. J.* 6, No. 2 (1963), 163-168.
- Fletcher R. and Reeves C. M. "Function minimization by conjugate gradients", *Comput. J.* 7, No. 2 (1964), 149-154.
- Frank M. and Wolfe P. "An algorithm for quadratic programming", *Nav. Res. Log. Quart.* 3, (1956), 95-110.

- Fridman V. M. "On the convergence of methods of steepest descent type", *Uspekhi Mat. Nauk* 17, No. 3 (1962) (in Russian).
- Gale D. *The Theory of Linear Economic Models*, New York, 1960.
- Gilbert E. G. "An iterative procedure for computing the minimum of a quadratic form on a convex set", *SIAM J. Contr.* 4, No. 1 (1966), 61-80.
- Glasman I. M. "Relaxation methods", in *Trudy pervoi zimnei shkoly po matem. progr.* (Drogobych, 1968), v. I, Moscow, 1969 (in Russian).
- Goldfarb D. "A family of variable-metric methods derived by variational means", *Math. Comput.* 24, No. 109 (1970), 23-26.
- Goldstein A. A. 1. "On steepest descent", *SIAM J. Contr.* 3 (1965), 147-151.
2. "Minimizing functionals on normed linear spaces", *SIAM J. Contr.* 4, No. 1 (1966), 81-89.
- Goldstein A. A. and Price J. F. "An effective algorithm for minimization", *Num. Math.* 10 (1967), 184-189.
- Golstein E. G. 1. "Dual problems of convex and partly-convex programming in functional spaces" in *Issledovaniya po matematicheskomu programmirovaniyu*, Moscow, 1968 (in Russian).
2. *Convex programming: Elementary theory*, Moscow, 1970 (in Russian).
- Greenstadt J. "Variations on variable-metric methods", *Math. Comput.* 24, No. 109 (1970), 1-22.
- Hayes R. M. "Iterative methods of solving linear problems in Hilbert space" in "Contributions to the solution of systems of linear equations and determination of eigenvalues", *Nat. Bur. Stand. Appl. Math.* 39 (1954), 71-104.
- Halkin H. and Neustadt L. W. "General necessary conditions for optimization problems", *Proc. Nat. Acad. Sci.* 56 (1956), 1066-1071.
- Hestenes M. R. 1. "Multiplier and gradient methods", *J. Optim. Theory Applic.* 4, No. 5 (1969).
2. *Calculus of Variations and Optimal Control Theory*, New York, 1966.
3. "The conjugate gradient method for solving linear systems", *Proc. Symp. Appl. Math.* 6 (1956), 86-102.
- Hestenes M. R. and Stiefel E. "Methods of conjugate gradients for solving linear equations", *J. Res. Nat. Bur. Stand.* 49, No. 6 (1952), 409-436.
- Horwitz L. B. and Sarachik P. E. "Davidon's method in Hilbert space", *SIAM J. Appl. Math.* 16, 4 (1968).
- Huang H. Y. "Unified approach to quadratically convergent algorithms for function minimization", *J. Opt. Theory Appl.* 5, No. 6 (1970).
- Huang N. Y. and Levy A. V. "Numerical experiments on quadratically convergent algorithms for function minimization", *J. Opt. Theory Appl.* 6, No. 3 (1970), 269-282.
- Huard P. "Resolution of mathematical programming with nonlinear constraints by the method of centers" in *Nonlinear Programming*, J. Abadie (ed.), Amsterdam, 1967, pp. 208-219.
- Ivanov V. V. 1. *Theory of Approximation Methods and Its Application to the Numerical Solution of Singular Integral Equations*, Kiev, 1968 (in Russian).
2. *On the Accuracy and Effectiveness of Computational Algorithms*, Kiev, 1969 (in Russian).
3. "On optimal algorithms of minimizing functions of certain classes", *Kibernetika*, No. 4 (1972), 81-94 (in Russian).
- Ivanov V. V. and Truten' V. E. "Analysis of the accuracy of quadratic programs", *Kibernetika*, No. 4 (1969), 94-105 (in Russian).
- Isaev V. K. and Sonin V. V. "On a modification of the Newton method for numerical solving of boundary value problems", *Zh. Vychisl. Mat. i Mat. Fiz.* 3, No. 6 (1963), 1114-1116 (in Russian).
- Kantorovich L. V. 1. "On the method of steepest descent", *Dokl. Akad. Nauk SSSR* 56 (1947), 233-236 (in Russian).
2. "On the Newton method for functional equations", *ibid.* 60, No. 7 (1948) (in Russian).

- Kantorovich L. V. and Akilov G. P. *Functional Analysis in Normed Spaces*, New York, 1964.
- Karlin S. *Mathematical Methods and Theory in Games, Programming and Economics*, Reading, Mass., 1962.
- Karmanov V. G. "Estimates of rate of convergence of iterative methods of minimization", *Zh. Vychisl. Mat. i Mat. Fiz.* 14, No. 1 (1974) (in Russian).
- Kelley H. J. "The gradient method" in *Optimization techniques*, New York, 1962.
- Kelley J. E. "The cutting plane method for solving convex programs", *J. Soc. Ind. Appl. Math.* 8, No. 4 (1960), 703-712.
- Kolmogorov A. N. and Fomin S. V. *Introductory Real Analysis*, Englewood Cliffs, N. J., 1970.
- Künzi H. and Krelle W. *Nichtlineare Programmierung*, Berlin, 1962.
- Künzi H. and Oettli W. *Nichtlineare Optimierung: neuere verfahren bibliographie*, Berlin-Heidelberg-New York, 1969.
- Laenberger D. "Convergence rate of penalty-function scheme", *J. Opt. Theory and Appl.* 7, No. 1 (1971), 39-51.
- Lavrov S. S. "Application of barycentric coordinates for solving some numerical problems", *Zh. Vychisl. Mat. i Mat. Fiz.* 4, No. 5 (1964), 905-911 (in Russian).
- Levitin E. S. and Polyak B. T. "Constrained minimization methods", *Zh. Vychisl. Mat. i Mat. Fiz.* 6, No. 5 (1966), 787-823 (in Russian).
- Lusternik L. A. and Sobolev V. I. *Elements of Functional Analysis*, New York, 1974.
- Lyubich Yu. I. "Steepest descent", *Trudy vtoroi zimnei shkoly po matem. progr. i smezhn. vopr.* Issue 1, Moscow, 1969, 113-151 (in Russian).
- Lyubich Yu. I. and Maistrovsky G. D. "General theory of relaxation processes for convex functionals", *Uspekhi Mat. Nauk* 25, Issue 1, 1970 (in Russian).
- Maistrovsky G. D. 1. "On the convergence of the conjugate gradient method", *Zh. Vychisl. Mat. i Mat. Fiz.* 11, No. 5 (1971), 1291-1294 (in Russian).
2. "Proof of quadratic convergence of the conjugate gradient method", *Vychisl. Mat. i Vychisl. Tekhn. Fiz.-techn. inst. nizkikh temperatur*, Kharkov, Issue 2 (1971), 3-5 (in Russian).
- McCormick G. P. and Pearson J. D. "Variable metric methods and unconstrained optimization", Confer. on Optimal., Keele Hall, England, March 1968.
- Miele A., Huang H. Y. and Heideman J. C. "Sequential gradient-restoration algorithm for the minimization of constrained function—ordinary and conjugate gradient versions", *J. Opt. Theory Appl.* 4, No. 4 (1969).
- Moiseev N. N. (ed.) *Numerical Methods in the Theory of Optimal Systems*, Moscow, 1971 (in Russian).
- Murtagh B. A. and Sargent R. W. H. "Computational experience with quadratically convergent minimization methods", *Comput. J.* 13, No. 2 (1970), 185-194.
- Neustadt L. W. "An abstract variational theory with applications to broad class of optimization problems. I. General theory", *SIAM J. Contr.* 4, No. 3 (1966), 505-527.
- Oblomskaya L. Ya. "A comparison of the rate of convergence of the conjugate gradient method and of the gradient method for quadratic functionals", in *Voprosy tochnosti i effektivnosti vychislitelnykh algoritmov* (Problems of Accuracy and Effectiveness of Computation Algorithms), 4 (1968), Kiev, 94-103 (in Russian).
- Ostrowski A. M. *Solution of Equations and Systems of Equations*, 2nd ed., New York, 1966.
- Pearson J. D. "Variable metric methods of minimization", *Comput. J.* 12, No. 2 (1969), 171-178.
- Polak E. 1. "On primal and dual methods of solving discrete optimal control problems" in *Computing Methods in Optimization Problems—2*, New York, 1969.

2. *Computational Methods in Optimization: a Unified Approach*, New York, 1971.
- Polyak B. T. 1. "Gradient methods for functional minimization", *Zh. Vychisl. Mat. i Mat. Fiz.* 3, No. 4 (1963), 643-654 (in Russian).
2. "Method of conjugate gradients", *Trudy vtoroi zimnei shkoly po matem. program. i smezhn. vopr.* Issue 1, Moscow, 1969, 152-201 (in Russian).
 3. "Iterative methods using Lagrange multipliers for solving extremal problems with equality constraints", *Zh. Vychisl. Mat. i Mat. Fiz.* 10, No. 5 (1970), 1098-1106 (in Russian).
 4. "On the rate of convergence of the penalty function method", *ibid.* 11, No. 1 (1971), 3-11 (in Russian).
- Powell M.J.D. 1. "A survey of numerical methods for unconstrained optimization", *SIAM Rev.* 12, No. 1, 1970, 79-97.
2. "An efficient method for finding the minimum of a function of several variables without calculating derivatives", *Comput. J.* 7, No. 2 (1964), 155-162.
 3. "On the Convergence of the Variable Metric Algorithm", Mathematics Branch, Atomic Energy Research Establishment, Harwell, Berkshire, England, October 1969 (mimeo).
- Pshenichny B. N. 1. *Necessary Conditions for an Extremum*, New York, 1974.
2. "The duality principle in problems of convex programming", *Zh. Vychisl. Mat. i Mat. Fiz.* 5, No. 1 (1965), 98-106 (in Russian).
 3. "On a descent algorithm", *ibid.* 8, No. 3 (1968), 649-652 (in Russian).
 4. "Algorithms for the general problem of mathematical programming", *Kibernetika*, No. 5 (1970), 120-125 (in Russian).
 5. "On the acceleration of convergence of algorithms for solving optimal control problems" in *Computing Methods in Optimization Problems*, New York, 1969.
- Pshenichny B. N. and Ganzhela I. F. "An algorithm for solving the problem of convex programming with linear constraints", *Kibernetika*, No. 3 (1970), 81-85 (in Russian).
- Rockafellar R. T. *Convex Analysis*, Princeton, N. J., 1970.
- Rosen J. B. "The gradient projection method for nonlinear programming. Part I: Linear constraints. Part II: Nonlinear constraints", *SIAM J. Appl. Math.* 8, (1960), pp. 181-217; 9 (1961), 514-532.
- Shamansky V. E. 1. *Methods of Numerical Solving of Boundary Value Problems on Computers*. Part II, Kiev, 1966 (in Russian).
2. "On some computation schemes of iterative processes", *Uspekhi Mat. Nauk*, 14, No. 1 (1962), 100-109 (in Russian).
- Shor N. Z. 1. "Generalized gradient descent", *Trudy pervoi zimnei shkoly po mat. progr.* Drogobych, Moscow, 1969, 578-585 (in Russian).
2. "On the rate of convergence of the generalized gradient descent", *Kibernetika*, No. 3 (1968), 98-99 (in Russian).
 3. "Using the operation of space stretching in problems of convex function minimization", *Kibernetika*, No. 1 (1970), 6-12 (in Russian).
 4. "On the rate of convergence of the generalized gradient method with space stretching", *Kibernetika*, No. 2 (1970), 80-85 (in Russian).
- Smith C. S. "The automatic computation of maximum likelihood estimates", N.C.B. Sci. Dept. Report SC 846 (MR) 40.
- Smolyak S. A. "Quadratic rate of convergence of the conjugate gradients method", *Trudy tret'ei zimnei shkoly po mat. progr.* Moscow Building Institute, Moscow, 1970.
- Sorensen H. W. "Comparison of some conjugate direction procedures for function minimization", J. Franklin Institute, 288, 421 (1969).
- Tikhonov A. N. 1. "Regularization of incorrectly posed problems", *Dokl. Akad. Nauk SSSR*, 153 (1963), 49-52 [Engl. trans.: *Soviet Math. Doklady* 4 (1963), 1624].

2. "On the stability of algorithms for solving degenerate systems of linear algebraic equations", *Zh. Vychisl. Mat. i Mat. Fiz.* 5, No. 4 (1965) (in Russian).
- Tokumaru H., Adachi N. and Goto K. "Davidon's method for minimization problems in Hilbert space with an application to control problems", *SIAM J. Contr.* 8, No. 2 (1970).
- Topkis D. M. and Veinott A., Jr. "On the convergence of some feasible directions algorithms for nonlinear programming", *SIAM J. Contr.* 5, No. 2 (1967), 268-279.
- Vainberg M. M. 1. *Variational Methods for the Study of Nonlinear Operators*, San Francisco, 1964.
2. *Variational Method and Method of Monotone Operators in the Theory of Nonlinear Equations*, New York, 1973.
- Vasil'ev F. P. *Lectures on the Methods for Solving Extrema Problems*, Moscow State University, 1974 (in Russian).
- Warga J. "A convergent procedure for convex programming", *J. Soc. Ind. and Appl. Math.* 11, No. 3 (1963), 579-587.
- Yakovlev M. N. "On some methods of solving nonlinear equations", *Trudy Matem. Inst. AN SSSR* 84 (1965), 8-40 (in Russian).
- Zangwill W. I. 1. "Minimizing a function without calculating derivatives", *Comput. J.* 10, No. 3 (1967), 293-296.
2. "Nonlinear programming via penalty functions", *Management Science* 13, No. 5 (1967), 344-368.
- Zelevnik F. J. "Quasi-Newton methods for nonlinear equations", *J. Assoc. Comput. Mach.* 15, No. 2 (1968), 265-271.
- Zoutendijk G. 1. *Methods of Feasible Directions*, Amsterdam, 1960.
2. "Nonlinear programming: A numerical survey", *SIAM J. Contr.* 4, No. 1 (1966), 194-210.
- Zukhovitskii S. I. and Avdeyeva L. I. *Linear and Convex Programming*, Philadelphia, Pa., 1966.
- Zukhovitskii S. I., Polyak R. A. and Primak M. E. 1. "An algorithm for the solution of the problem of convex Chebyshev approximation", *Doklady Akad. Nauk SSSR* 151, No. 1 (1963) 27-30 [Engl. trans.: *Soviet Math. Doklady* 4 (1963), 901].
2. "An algorithm for the solution of the convex programming problem", *Doklady Akad. Nauk SSSR* 153, No. 5, (1963), 991-994 [Engl. trans.: *Soviet Math. Doklady* 4 (1963), 1754].

INDEX

A

Abadie, J. 267
acceleration of convergence 224ff
 algorithm for 230
 and mathematical programming 232
Adachi, N. 270
admissible direction 33
admissible domain 32
Akilov, G. P. 268
algorithm(s)
 computational schemes of 259ff
 for conjugate directions method 93,
 111, 126, 129, 260
 for cutting hyperplane method 185
 for dual directions method 129, 259
 effectiveness of 60, 66
 for feasible directions method 261
 for linearization method 262
 for minimizing functions 139
 for quadratic programming 151,
 158, 160
Altman, M. 145, 265
Arrow, K. 43, 265
Auslender, A. 265
Avdeyeva, L. I. 42, 43, 257, 270

B

Balakrishnan, A. V. 265
Barnes, J. 258, 265
Bershchansky, Ya. M. 265
biorthogonalization, 125
boundary manifold, 38
Bourbaki, N. 23, 265
Brent, R. P. 145, 265
Brown, K. M. 265
Broyden, C. G. 128, 258, 265
Budak, B. M. 265

C

Canon, M. D. 176, 257, 265
Cauchy, A. L. 145, 265
Céa, J. 258, 265
closure of convex set 14
Collatz, L. 145, 265
concave function 23
conditional gradient method 170ff
 and step adjustment 177
cone
 conjugate 14
 convex 14
 polyhedral 15
conjugate directions 82
 and dual directions 143
conjugate directions methods 138, 143,
 145, 255
conjugate gradients method 98, 145-
 147, 152, 153, 156, 161
 algorithm for 260
conjugate vector 82, 138, 141, 144
 algorithms for construction 87
constrained function minimization
 146ff
constraints
 linear 146
 simple 160
convergence
 acceleration of 224ff
 for conditional gradient algorithms
 172, 176
 of conjugate directions methods
 104, 119, 125, 144
 of dual directions methods 144
 rates of 11
convex cone 14
convex functions 17ff
 in finite-dimensional space 42

convex functions
 in functional spaces 42
 convex programming 24ff
 and feasible directions 162
 Fiacco and McCormick method 243
 in finite-dimensional spaces 42
 and the penalty function method 240, 242
 quality of 43
 convex quadratic function
 minimization of 99
 convex sets 12ff
 in finite-dimensional space 42
 in functional spaces 42
 Cullum, C. D. 176, 257, 265
 cutting hyperplane method 184ff
 algorithm for 185
 computational aspects of 187
 for convex programming 185
 and the dual problem 187
 and linear programming 187
 rate of convergence of 188

D

Daniel, J. W. 145, 265
 Danilin, Yu. M. 145, 257, 258, 265, 266
 Dantzig, G. B. 43, 266
 Davidon, W. C. 128, 145, 266
 Dem'yanov, V. F. 266
 Dennis, I. E. 266
 Dennis J. E. 265
 direction feasibility method 162ff
 direction of motion 132
 dual directions
 and conjugate directions 143
 dual directions method(s) 130, 136, 143, 255
 algorithm for 259
 dual problem
 in convex programming 28, 29
 and the cutting hyperplane method 187
 in linear programming 30
 and the minimization problem 29
 and projection operators 160
 in quadratic programming 194, 195

dual systems of vectors 77
 Dubovitsky, A. Ya. 43, 266
 Dunford, N. 42, 266

E

Eaton, J. H. 265
 effectiveness
 of conjugate directions methods 128
 effectiveness of algorithm 66
 and the number of iterations 66
 effectiveness of methods
 and rate of convergence 66
 Eremin, I. I. 266
 Ermol'ev, Yu. M. 258, 266
 extrema
 necessary conditions for 42, 43

F

Faddeev, D. K. 76, 81, 94, 145, 159, 266
 Faddeeva, V. N. 76, 81, 94, 145, 159, 266
 feasible directions 162
 choice of 162
 and convex programming 162
 feasible directions method 162ff
 algorithm for 261
 feasible point 149
 Fiacco, A. V. 258, 266
 Fiacco and McCormick method 243
 Fikhtenhol'ts, G. M. 43, 266
 finite differences
 and gradient 129, 137
 first order methods 248, 249
 rate of convergence of 251
 Fletcher, R. 128, 145, 266
 Fomin, S. V. 42, 268
 Frank, M. 257, 266
 Fridman, V. N. 267
 function minimization 137, 139
 constrained 146ff
 unconstrained 44ff

G

Gale, D. 43, 267
 Ganzhela, I. F. 269
 generalized Newton's method 58ff

- p>geometrical-progression rate of convergence 11
-
- Gilbert, E. G. 267
-
- Glasman, I. M. 267
-
- Goldfarb, D. 267
-
- Goldman, N. R. 265
-
- Goldstein, A. A. 145, 267
-
- Golstein, E. G. 43, 267
-
- Goto, K. 270
-
- gradient
-
- and finite differences 129, 137
-
- gradient method(s) 45ff
-
- effectiveness of 56, 57
-
- qualitative analysis of 55
-
- rate of convergence of 145
-
- of steepest descent 45
-
- Greenstadt, J. 128, 267
-
- H
-
- Halkin, H. 43, 267
-
- Hayes, R. M. 267
-
- Heideman, J. C. 258, 268
-
- Hestenes, M. R. 43, 145, 258, 267
-
- Horwitz, L. B. 43, 265, 267
-
- Huang, H. Y. 128, 145, 258, 268
-
- Huang, N. Y. 267
-
- Huard, P. 258, 267
-
- I
-
- Isaev, V. K. 267
-
- Ivanov, V. V. 257, 258, 267
-
- K
-
- Kantorovich, L. V. 145, 267, 268
-
- Karlin, S. 42, 43, 268
-
- Karmanov, V. G. 258, 268
-
- Kelley, H. J. 268
-
- Kelley, J. E. 145, 188, 257, 268
-
- Kolmogorov, A. N. 42, 268
-
- Krelle, W. 42, 257, 268
-
- Kuhn-Tucker theorem 27, 196, 217, 218, 220
-
- Künzi, H. 42, 257, 258, 268
-
- L
-
- Laenberger, D. 258, 268
-
- Lagrange multipliers 28
-
- Lagrange's formula for operators 42
-
- Lagrange's generalization formula 42
-
- Lavrov, S. S. 268
-
- Levitin, E. S. 42, 188, 257, 258, 268
-
- Levy, A. V. 128, 267
-
- linear constraints 146
-
- linear independence of vectors 219
-
- linearization method 188ff
-
- acceleration of convergence for 233
-
- algorithm for 190, 262
-
- convergence of algorithm 190, 193, 194, 213
-
- and the dual problem 198
-
- and finding the minimax 211ff
-
- and linear programming 200
-
- and simplex method 200
-
- and systems of equalities and inequalities 211ff
-
- linearized objective function 188
-
- linear programming 29, 43
-
- algorithms for 43
-
- and cutting hyperplane method 187
-
- in finite dimensional spaces 42
-
- and necessary conditions for a minimum 201, 204
-
- and sufficient conditions for a minimum 204
-
- linear rate of convergence 11
-
- Lipschitz' condition 46
-
- local acceleration of convergence 224ff
-
- local minimum
-
- point of 33
-
- Lusternik, L. A. 245, 268
-
- Lyubich, Yu. I. 145, 268
-
- M
-
- Maistrovsky, G. D. 10, 145, 268
-
- Malozemov, V. N. 257, 266
-
- mathematical programming 12ff
-
- the general problem of 256
-
- matrices of second derivatives 66
-
- finite differences analogue 75
-
- McCormick, G. P. 258, 266, 268
-
- method of conditional gradient 170ff
-
- and Newton's method with step adjustment 177
-
- rate of convergence 172, 176

- methods of conditional gradient
 - method(s) of conjugate directions 82ff
 - algorithms for 93
 - applicability of 103
 - construction of 85
 - convergence of 104, 109, 125f
 - effectiveness of 84, 86, 128
 - and minimization of functions 103ff
 - and minimization of quadratic functions 82
 - properties of 89
 - properties of algorithms 111, 126
 - rate of convergence 144
 - method of conjugate gradients 98ff
 - algorithm for 260
 - method(s) of dual directions 67ff
 - algorithm for 74, 259
 - choice of schemes for 67, 68
 - construction of iterative processes 68
 - and minimization of quadratic functions 79, 80
 - and Newton's method 80, 81
 - properties of 80
 - rate of convergence 144
 - substantiation of 69
 - without calculating derivatives 145
 - method of feasible directions 162ff
 - algorithm for 165, 261
 - applied to linear programming 163, 166
 - convergence of 167, 170
 - methods of first order 248, 249
 - rate of convergence 251
 - methods of second order 246, 252
 - rate of convergence 254
 - method of steepest descent 45
 - method(s) without calculating derivatives 129ff
 - Miele, A. 258, 268
 - Milyutin, A. A. 43, 266
 - minimax problem 39
 - minimization
 - of a convex quadratic function 79, 99
 - and equality constraints 233
 - of nonlinear functions 170
 - of quadratic functions 82, 146, 148, 149
 - minimization of functions
 - methods of conjugate directions 103ff
 - minimizing quadratic functions
 - by method of dual directions 80
 - minimum of function
 - necessary conditions for 25, 32ff
 - second-order necessary conditions 40
 - minimum point 33
 - Moiseev, N. N. 10, 268
 - Murtagh, B. A. 268
- N
- necessary conditions for a minimum 180, 194
 - in linear programming 201
 - Neustadt, L. W. 43, 265, 267, 268
 - Newton-Leibnitz formula 41
 - Newton's finite difference method 81
 - Newton's method
 - and the conditional gradient method 177
 - and conjugate gradient 170ff
 - convergence of 215, 244
 - and dual directions 81
 - generalized 58ff
 - rate of convergence 178, 182
 - Newton's method with step adjustment 58ff
 - choosing of step 59
 - and gradient methods 66
 - and methods of dual directions 80
 - properties of 59-67
 - nondegeneracy assumption 164
 - nonlinear equality constraints
 - and the linearization method 188
 - nonlinear functions
 - minimization of 170
- O
- objective function 33
 - linearized 188

Oblomskaya, L. Ya. 268
 Oettli, W. 258, 268
 Ostrowski, A. M. 232, 268
 Ostrowski's theorem 206, 210, 211

P

Pearson, J. D. 128, 268
 penalty function method 235ff
 computational aspects 242
 in convex programming 240
 substantiation of 236
 point of local minimum 33
 Polak, E. 268, 269
 Polyak, B. T. 42, 128, 145, 188, 257, 258, 268, 269
 Polyak, R. A. 270
 polyhedral cone 15
 Powell, M. J. D. 128, 145, 266, 269
 Price, J. F. 145, 267
 Primak, M. E. 257, 270
 primal problem in linear programming 30
 projection methods with restoration of ties 244ff
 projection operator 147, 209
 and the dual problem 160
 matrix for 158
 Pshenichny, B. N. 43, 128, 145, 257, 258, 266, 269

Q

quadratic forms
 minimization of 137, 143
 quadratic functions
 minimization of 146, 148, 149
 quadratic programming 31, 146ff
 algorithm for 151, 158, 160
 and the dual problem 194, 195
 iterative process for 157
 with simple constraints 160
 quadratic rate of convergence

R

rate of convergence
 for conditional gradient algorithms 172, 176

and effectiveness of method 60
 of first order methods 251
 geometrical-progression 11
 linear 11
 of methods of conjugate directions 119, 125f
 quadratic 11
 of second order methods 254
 superlinear 11

Reeves, C. M. 145, 266
 regular point 37
 restoration process 104, 105
 Rockafellar, R. T. 42, 43, 269
 Rosen, J. B. 258, 269
 Rubinov, A. M. 257, 266

S

Sarachik, P. E. 267
 Sargent, R. W. H. 268
 Schwartz, J. T. 42, 266
 second order methods 246, 252
 rate of convergence of 254
 separation theorem 12
 Shamansky, V. E. 258, 269
 Shor, N. Z. 257, 258, 269
 simple constraints 160
 simplex method
 and the linearization method 200
 in solving the dual problem 187
 Slater's condition 217, 220
 Smith, C. S. 145, 269
 Smolyak, S. A. 145, 269
 Sobolev, V. I. 245, 268
 Sonin, V. V. 267
 Sorensen, H. W. 269
 step length
 algorithm for choosing 132
 choice of 171, 172
 Stiefel, E. 145, 267
 strictly convex functions 21, 42
 strictly convex sets 17
 strongly convex functions 21, 42
 strongly convex sets 17
 strongly positive matrix 22
 subgradient 20

sufficient condition

for a local minimum 234

superlinear rate of convergence 11

support vector 20, 186

system(s) of conjugate vectors 141, 144

T

tangent manifold 38

Tikhonov, A. N. 10, 269, 270

Tokumaru, H. 270

Topkis, D. M. 257, 270

Truten', V. E. 257, 267

U

unconstrained function minimization
44ff

Uzawa, H. 43, 265

V

Vainberg, M. M. 42, 270

Vasil'ev, F. P. 258, 270

vector of dual variables 29

Veinott, A. 257, 270

W

Warga, J. 270

Weierstrass' theorem 24, 178, 249

Wolfe, P. 257, 266

Y

Yakovlev, M. N. 145, 270

Z

Zangwill, W. I. 145, 258, 270 .

Zelevnik, F. J. 258, 270

Zoutendijk, G. 42, 257, 258, 270

Zukhovitskii, S. I. 42, 43, 257, 270

To the Reader

Mir Publishers would be grateful for your comments on the content, translation and design of this book. We would also be pleased to receive any other suggestions you may wish to make.

Our address is:

USSR, 129820, Moscow I-110, GSP

Pervy Rizhsky Pereulok 2

Mir Publishers

Printed in the Union of Soviet Socialist Republics

OTHER MIR TITLES

N. S. Bakhvalov

Numerical Methods

Professor N. S. Bakhvalov of Moscow State University has written a textbook for advanced courses in applied mathematics. It is designed especially for universities and higher technical schools and is concerned with the theory and practice of numerical methods. The text covers a wide range of subjects including the essentials of numerical methods dealing with approximation of functions, integration, problems of algebra and optimization, and the solution of ordinary differential equations. Particular attention is paid to questions involving choice of method and organization of computations in the solution of large sets of problems of a single type. The book includes a bibliography of Russian and English references.

V. I. Krylov and N. S. Skoblya

*A Handbook of Methods of Approximate
Fourier Transformation and Inversion
of the Laplace Transformation*

Harmonic analysis and Laplace transformations are very often used in the solution of many theoretical and practical problems. This text contains most of the known methods of the approximate inversion of the Laplace transformation and methods of calculating Fourier integrals. It is designed for scientists and engineers who have to do with the theory of applications of the Laplace transformation and the Fourier integrals. It can serve as a useful reference work for computing centres and design bureaus.

V. P. Maslov

Operational Methods

This is a textbook for second- and third-year university mathematics and physics students based on the author's lectures at the faculty of applied mathematics of the Moscow Institute of Electronic Engineering and the physics faculty of Moscow University. It illustrates the theoretical material concerning specific physical problems, which are taken as the model, comparing the formulas of the operational method with the numerical solution. The book will be of interest to scientific workers in general.